Published in final edited form as:
Ann Epidemiol. 2016 October ; 26(10): 674–680. doi:10.1016/j.annepidem.2016.08.016.

Does water kill? A call for less casual causal inferences

Miguel A. Hernán[1,2]

[1]Departments of Epidemiology and Biostatistics, Harvard T.H. Chan School of Public Health, Boston, MA 02115

[2]Harvard-MIT Division of Health Sciences and Technology, Boston, MA 02115

Abstract

"Can this number be interpreted as a causal effect?" is a key question for scientists and decision makers. The potential outcomes approach, a quantitative counterfactual theory, describes conditions under which the question can be answered affirmatively. This article reviews one of those conditions, known as consistency, and its implications for real world decisions.

Introduction

Long gone are the times when causality was the exclusive realm of philosophers and theologians. Today's scientists embrace causal inference explicitly as a legitimate endeavor. To make causal inferences, scientists do what they do best: they generate numbers. To understand the conditions under which those numbers can be interpreted as causal effects, scientists use a quantitative counterfactual theory which is often referred to as the potential outcomes approach. (This paper will treat "quantitative counterfactual theory" and "potential outcomes approach" as equivalent terms; finer distinctions between them may be proposed but are of little relevance for our discussion.) This theory was formalized by Neyman[1] for randomized experiments, extended to nonexperimental—or observational—studies with time-fixed treatments by Rubin,[2,3] and generalized to randomized and observational settings with time-varying treatments by Robins.[4,5]

The potential outcomes approach provides conceptual definitions and supports analytic methods for researchers interested in producing and interpreting numerical estimates of causal effects. However, the potential outcomes approach is not universally accepted. In this issue of the journal, Schwartz and collaborators[6] criticize some restrictions imposed on causal inference by quantitative counterfactual theory. A recent article warns that the potential outcomes approach is damaging and "hardline", is based on "a restrictive set of convictions", makes "imperious claims", and "cannot explain how [other approaches work]".[7] Others have said that the potential outcomes approach is "socially conservative", because "it neglects, discourages, and dismisses […] radical change."[8] The critics of quantitative counterfactual theory make it look narrow-minded, arrogant, and even reactionary.

This commentary is based on a talk delivered at the John Snow Lecture Theatre, London School of Hygiene and Tropical Medicine, during the UK Causal Inference Meeting, April 14th, 2016

This paper is an attempt to address those criticisms, and to clarify three common misunderstandings. Several colleagues have recently written lucid commentaries that, with different emphases, address these issues too.[9–12]

As we will see, a first misunderstanding occurs because the potential outcomes approach explicitly highlights the inherent vagueness of all causal questions. In a perfect example of a shoot-the-messenger attitude, some critics have used this transparency of the approach as ammunition against it.

A second misunderstanding arises when critics ask too much from quantitative counterfactual theory. The potential outcomes approach is concerned with questions of the sort "what is the average causal effect of A on Y?" in a particular setting. The answer to these questions is a number (or several numbers) plus a statistical measure of uncertainty. In contrast, philosophical discussions about causality often revolve about questions of the sort "is A a cause of Y?" The answer to these questions is 'yes' or 'no', not a number. The potential outcomes approach does not privilege a particular definition of "cause" and therefore may not necessarily provide definite yes/no answers to questions about causes.[11]

A third misunderstanding is the idea that the potential outcomes framework restricts causal inference to the effects of humanly feasible, or practicable, interventions.[13] This is not necessarily the case. Much of the trepidation about the potential outcomes approach dissolves after clarifying that the framework is not restricted to feasible interventions.

To help clarify these misunderstandings, we first review the definition of causal contrast and its reliance on a fundamental condition for causal inference from observational data: consistency. Then we dissect the components of the consistency condition and their implications for estimating causal effects. After briefly discussing the role of two other commonly used conditions (exchangeability and positivity) for causal inference from observational data, we propose a taxonomy of causal questions and their political implications. The goal is to clarify the role of a quantitative counterfactual theory for causal inference in both scientific and policy settings.

Causal contrasts

Questions about the causal effect of a treatment A on an outcome Y in a particular population can be expressed in terms of counterfactual contrasts. For example, we say that the average causal effect of the binary treatment A on the outcome Y is $\mathrm{E}[Y^{a=1}] - \mathrm{E}[Y^{a=0}]$, where $Y^{a=1}$ is the (counterfactual or potential) outcome that would have been observed if an individual had received treatment level a=1, $Y^{a=0}$ is the outcome that would have been observed if an individual had received treatment level a=0, and $\mathrm{E}[Y^a]$ is the mean of Y^a in the population of interest. For simplicity, we will assume that the population and the outcome Y of interest are well defined, and that no measurement error and no interference[14,15] exist.

The potential outcomes approach is used to estimate the numerical value of average causal effects like $\mathrm{E}[Y^{a=1}] - \mathrm{E}[Y^{a=0}]$. A non-zero average causal effect $\mathrm{E}[Y^{a=1}] - \mathrm{E}[Y^{a=0}] \neq 0$ can be viewed as a sufficient, but not necessary, condition for A to be "a cause".[11] Therefore, a counterfactual contrast like E[$Y^{a=1}$] − E[$Y^{a=0}$] can be used to define causal effects, but it cannot generally be used to identify causes because it cannot encompass all definitions of the term “cause.”

Suppose that we have identified a population (e.g., London residents in 1854) and that we are interested in estimating the average causal effect of exposure to water A on their risk of death Y over the next week. Specifically, say we are interested in whether exposure to water has a non-null average causal effect on their risk of death. That is, we ask the question “does water kill”? This question is, of course, problematic. What do we mean by exposure to water? A tsunami, a jump in the Thames, a hot bath, a water injection, drinking water from a particular source?

The inadequacy of the “does water kill?” question is even more obvious when we try to interpret the meaning of the counterfactual outcome $Y^{a=1}$: an individual’s 1-week survival under exposure to water. Again, which form of exposure to water? The counterfactual outcome is ill-defined because it can mean too many different things, and therefore the average causal effect E[$Y^{a=1}$] − E[$Y^{a=0}$] is ill-defined too. For the same reasons, the causal diagram A → Y is uninterpretable. The interpretation of counterfactual contrasts requires a relatively unambiguous definition of the meaning of a=1 and a=0, a requirement referred by some of us as “well-defined interventions”.

Yet a couple of recent commentaries[6,7] express serious reservations about the emphasis on “well-defined interventions” a in observational research. For them, this emphasis is sometimes misguided, if not outright reprehensible. In order to analyze the sources of this disagreement, let us start with what we agree on: the question “does water A kill?” is very vague because the potential outcomes ($Y^{a=1}$, $Y^{a=0}$) are very ill-defined. Alternatively, we can say that consistency, a key condition for the identifiability of causal effects from observational data, does not even approximately hold.

The consistency condition has been widely discussed in the statistical and epidemiologic literature.[16–19] In its simplest form, it is stated as $Y^a = Y$ for every individual with $A=a$. That is, among individuals who received treatment level $A=a$, their potential outcome Y^a under treatment level a is equal to their observed outcome Y. The apparent simplicity of the consistency condition is deceptive. In the next two sections, we unpack consistency by explicitly describing its two main components:[3] (1) the definition of the interventions a, and (2) their linkage to the data.

Component #1 of consistency: Sufficiently well-defined interventions

Let us start by discussing the causal effect of water on death. We will then extend our discussion to more pressing questions about the causal effect of factors like obesity and high blood pressure.

The treatment “water”

Consider the question “does exposure to water have a non-null average causal effect on the 1-week risk of death among London residents in 1854”? or, for brevity, “does water kill?” in this population. Since we agree this question is too vague, let us make it more precise. Say that we are not considering drowning by immersion or trauma by a flash flood, but rather ingesting water through one's mouth. Thus the question can be rephrased as

- Version #2. Does drinking water kill?

  We mean fresh water, not salty water.

- Version #3. Does drinking fresh water kill?

  How much water? 1000 liters per day will kill you.

- Version #4. Does drinking a swig of fresh water kill?

  What is the source of the water? Tap, fountain, directly from the river…

- Version #5. Does drinking a swig of water from the Broad Street pump kill?

  Over which period?

- Version #6. Does drinking a swig of water from the Broad Street pump between August 31 and September 10 kill?

  Ok, compared with what? With drinking 3 liters of beer?

- Version #7. Does drinking a swig of water from the Broad Street pump between August 31 and September 10 kill compared with drinking all your water from other pumps?

  What about other factors that may affect the causal effect of interest?

- Version #8. Does drinking a swig of water from the Broad Street pump between August 31 and September 10 and not initiating a rehydration treatment if diarrhea starts kill, compared with drinking all your water from other pumps?

And so on. The process of precisely specifying *a* never ends. Version #1,000,000 would be very long but still imprecise. It is impossible to provide an absolutely precise definition of a version of treatment.[19,20]

On the one hand, this is problematic because, when there are multiple versions of *a* and different versions lead to different outcomes, causal effects are not well-defined. On the other hand, the problem is not as serious as it seems because absolute precision in the definition of the versions of treatment is not needed. Further specification of versions of treatment is required only until no meaningful vagueness remains.[17,20] For example, scientists agree that the speed of the water from the pump, or the color of the clothes worn while drinking the water, need not be specified. Those aspects are deemed to be irrelevant because varying them would not lead to different mortality outcomes. That is, we only need s*ufficiently well-defined* versions of treatment *a* for a meaningful interpretation of both the potential outcomes Y^a and of the counterfactual contrasts that involve them.

Which begs the question of "How do we know that a version of treatment is sufficiently well-defined" or, equivalently, that no meaningful vagueness remains? The answer is "We don't". Declaring a version of treatment sufficiently well-defined is a matter of agreement among experts based on the available substantive knowledge. Today we agree that the speed of the water coming out of the pump is irrelevant for the risk of death from drinking it, but future research might prove us wrong if it is demonstrated that high-speed water alters the virulence of the cholera bacterium.

This is an intrinsic feature of causal inference: the articulation of causal questions is contingent on domain expertise and informal judgment. What we view as a scientifically meaningful, relatively well-defined causal question at present may turn out to be viewed as too vague in the future after learning that finer components of the versions of *a* affect the outcome and therefore the magnitude of the causal effect. The best we can do is to specify the versions of treatment with as much detail as we believe necessary, which is precisely what the protocols of randomized experiments do.

In fact, when analyzing observational data, we can imagine the hypothetical randomized experiment that we would have liked to conduct—the target experiment or target trial—to estimate the causal effect of interest. Specifying the protocol of the target trial[21–23] is a device to assist us to ask sharper causal questions. The versions of treatment that we are interested in become the interventions that would be implemented in the target trial (hence the use of the expression "well-defined interventions" interchangeably with "well-defined versions of treatment"). In principle, the interventions do not need to be humanly feasible or practicable, only sufficiently well-defined (more about this later). The anchoring of causal inferences to target trials has been more or less explicitly suggested by many authors, including Cochran,[24] Rubin,[3] Feinstein,[25] Dawid,[26] and Robins.[4,5]

Despite the above disclaimers being explicitly made in previous papers on this topic,[19,20,27] some critics have chosen to argue that the condition of well-defined interventions must be incorrect because no intervention can be perfectly well-defined.[7] This is the first misunderstanding mentioned in the Introduction. Obviously, if perfectly well-defined interventions were necessary, science would be paralyzed. Causal inference from either randomized experiments or observational data would be impossible. Imagine specifying, with absolute precision, how you would inject certain chemical into a mouse to quantify the effect of the chemical on mouse cancer. Such absolute precision would require that one specify the angle of the needle, the brand of the needle, the temperature of the needle, the force exerted by your thumb, whether you use the right or left thumb... Again, only "sufficiently well-defined interventions" are required, where the determination of an intervention being sufficiently well-defined is a matter of expert agreement.

Let us summarize what we have concluded about the question "does water kill?":

1. The original question was hopelessly vague because there are many possible versions of the treatment "water".
2. Certain degree of vagueness remains despite our efforts to refine the versions of treatment. However, experts are more comfortable with version #8 than with versions #1 and #2.

3. Experts believe that many versions of treatment are irrelevant (for example, the speed at which water flowed) and therefore they do not bother to specify them.

The “effect of exposure to water $A=a$” is a silly expression that leads to an uninterpretable potential outcome Y^a. The example was chosen precisely because its silliness is uncontroversial (and as an homage to John Snow in the London lecture hall named after him). The next step is to argue that what we have concluded about “the effect of exposure to water” applies to many other treatments or exposures with multiple versions.

Compound treatments with multiple versions

Consider the question “does obesity have a non-null average causal effect on the 10-year risk of death among United States residents in 2016?” or, for brevity, “does obesity shorten life?”[28] in this population. Say that we define obesity as having a body mass index greater than or equal to 30 kg/m^2 (you can replace BMI by your favorite measure of adiposity) and consider the following refinements to the question:

- Version #2. Does having a BMI≥30 shorten life?
- Version #3. Does having a BMI≥30 between ages 40 and 50 shorten life?
- Version #4. Does attaining a BMI=30 at age 40 and keeping a BMI≥30 between ages 40 and 50 shorten life?
- Version #5. Does having a BMI=24 at age 18, progressively gaining weight until attaining a BMI=30 at age 40 and keeping a BMI≥30 between ages 40 and 50 shorten life?
- Version #6. Does having a BMI=24 at age 18, progressively gaining weight until attaining a BMI=30 at age 40 and keeping a BMI≥30 between ages 40 and 50 shorten life compared with the same lifetime course until age 40 followed by progressive weight loss until attaining BMI=24 at age 50?

And so on. Each new version imposes additional restrictions on the specification of the target trial, but an absolutely precise specification is impossible. In fact, the above list of versions only scratches the surface. We did not specify, for example, *how* body weight is modified. This is a fundamental component of the definition of the treatment a in a randomized experiment because the value of the outcomes Y^a may vary depending on whether weight changes are achieved via interventions on diet, exercise, genes, microbiota, cigarette smoking, amputation, bariatric surgery, etc.[28] Again, anchoring causal inferences to a target trial helps sharpen the specification of the interventions of interest in observational analyses.

A sufficiently well-defined intervention needs to specify the start and end of the intervention, and the implementation of its different components over time. Each version in the list above is the result of making a particular decision about those aspects of the intervention (which are key to the specification of the target trial). Different scientists may disagree about these decisions. If our colleagues are interested in the effect of weight changes throughout the entire lifetime as opposed to between ages 40 and 50 only, they will propose different versions and, when conducting or emulating their target trial, they may obtain numerical estimates of causal effect that differ from ours. When the target trial is not specified, these numerical differences may muddle scientific discussions about replication of results and triangulation of evidence.

Let us continue this line of reasoning. If we provide version #2 to 100 scientists and ask them to design a target trial, we may end up with 100 very different trials and 100 very different effect estimates.[29] In contrast, if we provide version #6 (or, better, version #45) to 100 scientists, we would expect more similar target trials and effect estimates. The more precisely we define the versions, the fewer opportunities for miscommunication among scientists exist when numerical estimates do not agree.

At this point some readers may rightly note that the process of specifying the interventions has altered the original question. Version #6 is not about the causal effect of obesity any more, but about the causal effect of somehow losing weight between ages 40 and 50. Had we specified that the weight loss is achieved through severe caloric restriction, then we would be talking about the causal effect of caloric restriction between ages 40 and 50 rather than the causal effect of obesity. In this example, the more we focus on giving a relatively well-defined causal interpretation to our numerical estimates, the farther from the original question we seem to get.

But that is a good thing. Forcing us to refine the causal question, until it is agreed that no meaningful vagueness remains, is an essential contribution of quantitative counterfactual theory to science. In our previous example, we moved from the hopelessly vague "causal effect of water on death" to the better defined "causal effect of drinking a swig water from the Broad Street pump on death," which is different from the "causal effect of drinking a swig of seawater on death." We similarly move from the "causal effect of obesity on death" to the "causal effect of weight changes via caloric restriction between ages 40 and 50 years on death," which is different from the "causal effect of weight changes via heavy cigarette smoking between ages 20 and 60."

This discussion again highlights the reliance of the term "sufficiently well-defined interventions" on expert consensus, which by definition changes over time. During many centuries, the existence of a causal effect of yellow bile on health appeared self-evident for most physicians; today we do not even believe that yellow bile exists. A few decades after his time, John Snow would have asked his causal question about pump water in terms of the *Vibrio cholerae* bacterium. A few decades from now, scientists will probably refine our obesity question in terms of cellular modifications which we barely understand at this time.

Our discussion applies to any compound treatment A with multiple relevant versions, like obesity and high blood pressure. Rather than presupposing that "the average causal effect of A on death" is a well-defined quantity, the potential outcomes approach encourages us to formulate questions in such a way that the numerical effect estimates can be causally interpreted.

Yet the rejection of hopelessly ill-defined causal questions by the quantitative counterfactual theory is troubling for some people, as some recently published criticisms show. To these authors, it is self-evident that obesity may causally affect mortality, possibly because they believe in the existence of some hypothetical intervention on obesity that does not depend on its timing or mechanism, even if they cannot specify the actual nature of the intervention. Hence these authors view the potential outcomes approach as a straightjacket that prevents them from declaring that obesity is a cause. This is the second misunderstanding mentioned in the Introduction.

The potential outcomes approach was not designed to determine whether A is or is not a cause, but to quantify the magnitude of the causal effect of A on Y. This quantification is only possible when the interventions are sufficiently well-defined as argued above. In the absence of sufficiently well-defined interventions, the potential outcome approach is agnostic about causality. That is, the potential outcomes approach does not prohibit us from labeling obesity as a cause of death; it just reminds us that any numerical estimate we obtain for the average causal effect of obesity on death will be hard to interpret.

These interpretability problems are exacerbated in scientific disciplines, including epidemiology, that use observational analyses to handle complex causal questions for which precisely defined contrasts are not available yet.

But the specification of sufficiently well-defined interventions only ensures the first component of consistency. A relatively unambiguous interpretation of numerical estimates also requires the second component of consistency.

Component #2 of consistency: Linkage between interventions and the data

Suppose the interventions of interest are sufficiently well-defined. We can now proceed to conduct the target trial that implements those interventions or, if that is not feasible, to emulate the target trial using observational data. To do so, we need to have data with versions of treatment that correspond to the interventions of interest. As an extreme example, if we had prospective data from a human population exposed to water in varying degrees during a tsunami, we could not reasonably use those data to estimate the effect of exposure to water taken from the Broad Street pump.

Analogously, suppose we have observational data from a population of 50-year old humans in which some individuals became obese because of interactions between their genes, gut bacteria, and environmental exposures throughout their lifetimes. We do not expect that the 10-year risk of death in individuals who became obese through these mechanisms equals the 10-year risk of death of individuals who became obese through a decade of smoking cessation and physical inactivity. Therefore, we would not use those data to quantify the effect of hypothetical public health interventions on smoking and physical activity.

This is all about the equal sign in the consistency condition $Y^a = Y$. Sufficiently well-defined interventions a allow us to interpret the potential outcome Y^a, but not necessarily to obtain an effect estimate from a dataset in which the existing versions of treatment cannot be linked to the interventions a. In that case, an individual's potential outcome Y^a is not guaranteed to equal her observed outcome Y. Ill-defined interventions like “exposure to water”, “obesity”, “high blood pressure”, etc. do not allow us to interpret causal effect estimates, and neither do well-defined interventions in the absence of appropriate data. But unless we sufficiently define the interventions, we will be unable to detect a mismatch between the interventions of interest and the data at hand. Previous articles offer a more rigorous discussion of causal inferences concerning compound treatments with multiple versions.[19,30]

One way out of this problem is to assume that the effects of all versions of treatment are identical or at least all in the same direction. In some cases, this may be a reasonable assumption. For example, if interested in the average causal effect of high vs. normal blood pressure on stroke, empirical evidence suggests that lowering blood pressure through different pharmacological mechanisms results in similar outcomes. We might then argue that a precise definition of the versions of the treatment “blood pressure”, as well as a detailed characterization of the versions present in the data, is unnecessary in order to link the potential and observed outcomes. In other cases, however, the assumption seems to go against the available evidence. For example, if interested in the average causal effect of weight maintenance on death, empirical evidence suggests that some interventions would increase the risk (e.g., continuation of smoking) while others would decrease it (e.g., moderate exercise).

Any scientific discussion about whether all or some versions of treatment lead to the same causal conclusion rests, again, on expert consensus and judgement. Because experts are fallible, the best we can do is to make these discussions—and our assumptions—as transparent as possible, so that others can directly challenge our arguments.

Exchangeability and positivity

There is an additional reason why it is important to define and identify the versions of treatment when estimating average causal effects by emulating a target trial from an observational dataset: the versions of treatment are not randomly assigned and each of them may be partly determined by different factors. For example, caloric intake is affected by physical activity level, which has a direct effect on mortality and is therefore a confounder for the effect of caloric intake on mortality. Therefore, to identify the average causal effect on death of interventions to reduce obesity via different levels of caloric restriction, we would need to adjust for physical activity. And to identify the average causal effect on death of interventions to reduce obesity via different levels of physical activity, one would need to adjust for confounders for the effect of physical activity, which will generally differ from the confounders for caloric intake. The set of confounders required to achieve conditional exchangeability depends on the intervention and outcome of interest,[28] and the same goes for the set of confounders over which positivity is required. The formal expression of the conditions of exchangeability and positivity (sometimes jointly referred to as ignorability) under multiple versions of treatment are similar to that of the conditions for the identification of direct effects.[30]

We need to know the interventions of interest for a successful emulation of a target trial that intervenes on them, that is, to adjust for confounding. Suppose we consider the ill-defined treatment “obesity” and we compare the 10-year death risk of obese vs. nonobese individuals conditional on their history of physical activity, diet, and cigarette smoking. Then the estimated association corresponds to the effect of other versions of treatment, such as genes, microbiota, and environmental exposures (or to measurement error in the adjustment variables). Because these other versions of treatment may not be directly manipulable, estimating their effect is not of primary interest for those who need to make decisions about clinical or policy interventions that are available at this time. This observation leads naturally to an exploration of the third misunderstanding mentioned in the Introduction.

A taxonomy of well-defined interventions and their political implications

Suppose our goal is to estimate the average causal effect of a treatment A on an outcome Y using observational data. Because the interpretation of numerical estimates of the average causal effect requires sufficiently well-defined interventions, we carefully specify the interventions a that define the potential outcome Y^a. We can now classify these interventions as either (a) absent in the data or (b) present in the data.

If the interventions of interest are absent in the data, then our effect estimates cannot be directly linked to the causal questions of interest. Achieving that linkage requires that one bridge the empirical void between the potential and observed outcomes with theoretical knowledge.[31] An extreme option is to assume no direct effects or treatment-variation irrelevance[17]: the effect of the versions of treatment we are interested in is the same as the effect of the versions that actually occur in the data. A less extreme option is to combine the existing observational data with parametric assumptions to build a dynamic, mechanistic model for causal inference.

Of course, even when the interventions can be linked to the data, causal inference generally involves the combination of observational data with the untestable assumptions of exchangeability and positivity. But the lack of a linkage between the interventions of interest and the treatment versions requires qualitatively different assumptions about the mechanisms that bring about the outcome in the real-world. These additional assumptions can be encoded, for example, into the parameters of agent-based models or microsimulation models. These models are designed to compare the distribution of potential outcomes under various interventions and therefore fall squarely within the quantitative counterfactual theory framework.[32,33]

The interventions a can also be classified as (1) infeasible or (2) feasible. An example of sufficiently well-defined, but infeasible, interventions arises when considering the causal effect of a genotype. The implicit intervention is inserting or replacing a sequence of DNA at conception without altering any other components of the embryo. Because this intervention approximately corresponds to what happens at conception (perhaps conditional on the composition of neighboring pieces of DNA), it can be argued that the intervention is approximately present in the data. Therefore, we say that these genetic interventions are sufficiently well-defined and we feel comfortable providing a causal interpretation for the numerical estimates of effect. That is, because these genetic variations are present in the data, we can empirically obtain those interpretable numerical estimates under exchangeability and positivity.

Though sufficiently well-defined interventions do not need to be feasible or practicable for causal effects to be numerically interpretable, the determination of feasibility needs to be constantly updated. For example, genetic engineering of the sort implied by the above genetic interventions has been traditionally impossible to implement, but new genome editing tools (e.g., CRISPR-Cas9) may soon render obsolete the characterization of genetic interventions as technically infeasible. (Ethical constraints for genome editing will remain, but that is another story.)

Infeasible interventions, by definition, come with two problems. First, causal predictions about infeasible interventions cannot yet be verified by conducting a randomized experiment. Second, infeasible interventions do not necessarily translate into actionable causal inferences because the interventions cannot yet be implemented in practice.[34]

Therefore, if our goal were to bring about urgent change in the world, estimating the causal effect of infeasible interventions would not be the optimal approach. When we adopt the perspective of a decision-maker (e.g., physician, patient, public health officer, policy-maker, political activist) who needs to make a decision now based on the available observational data, rather than that of a philosopher interested in causality or a scientist interested in eventually explaining the world, we would prefer to focus on estimating the effect of feasible interventions that can be actually implemented to solve real world problems.

The tension between infeasible interventions and the urgency of action is illustrated by recent discussions about the role of race in the United States, where the empirical evidence on racial disparities with respect to health outcomes (e.g., life expectancy) is overwhelming. These white-black numerical discrepancies have been used to support the statement “race is a cause” of health outcomes.[35] The potential outcomes framework has little to say about that statement because the interventions of interest are extremely ill-defined. In fact, there is not even agreement with respect to the definition of “race”, which is often simplistically reduced to genetic differences.

On the other hand, the potential outcomes framework is the ideal tool to frame causal discussions about how to eliminate racial disparities.[36,37] By forcing investigators to carefully define real-world interventions, the potential outcomes approach takes the attention away from philosophical debates about the definition of cause towards the consideration of potentially implementable policies that can alleviate poverty, reduce discrimination, and increase economic opportunities for disadvantaged populations. If these interventions are relatively well-defined (e.g., raise taxes on wealthy individuals and large corporations, and redistribute the funds to health, education, pensions, and job creation during the next 3 generations) but cannot be linked to the versions of treatment in the data, then the causal effects will need to be estimated by combining existing data with detailed parametric models that summarize expert knowledge.

A key point is that a quantitative counterfactual theory is agnostic towards structural, even radical, change. Investigators who elect to study hypothetical interventions that can be directly linked to the data will typically end up asking narrower questions (e.g., the effects of continuing cigarette smoking and of affirmative action) than those who elect to study hypothetical interventions that cannot be easily linked to the data (e.g., the effects of increased taxation of tobacco products and regulation of manufacturers; and of high-quality education and health care for all families).[38] But all of these investigators are asking causal questions that can be represented as counterfactual contrasts and encapsulated into the protocol of a hypothetical experiment.

Conclusions

The goal of the potential outcomes framework is not to identify causes—or to "prove causality", as it sometimes said. That causality cannot be proven was already forcibly argued by Hume in the 18th century.[39] Rather, quantitative counterfactual inference helps us predict what would happen under different interventions, which requires our commitment to define the interventions of interest. As Rubin said in 1978: "Without treatment definitions that specify actions to be performed on experimental units, we cannot unambiguously define causal effects of treatments." The goal of the potential outcomes framework is to make quantitative causal inferences less casual.

Accusing the potential outcomes approach of not declaring a variable "a cause" is like accusing nuclear chemistry of not proving the existence of God. Those interested in the discovery of causes, as opposed to the interpretation of numerical estimates of causal effects, need to look for an alternative theoretical framework to the potential outcomes approach. Unfortunately, no such framework seems to be available. Recent calls for a "pluralistic" approach to causal inference resorted to philosophical arguments but did not propose any formal framework for the interpretation of causal effects.[7] Other approaches to quantitatively interpret effect estimates, like Dawid's decision-analytic perspective,[26,40,41] are consistent with the potential outcomes approach described here.[20]

The potential outcomes approach provides a vehicle for rigorous quantitative causal inference. This vehicle helps researchers stay on the paths opened by their predecessors who fought hard for evidence-based approaches —in which empirical findings may trump unconstrained expert opinion— while giving them the flexibility to steer this vehicle in the direction that their creativity suggests—for example, by articulating counterfactual contrasts based on novel negative controls that provide information about the direction or magnitude of a causal effect.[42]

The inherent vagueness of all causal questions is brought to the forefront by the potential outcomes approach. Defining the interventions of interest as well as possible helps reduce that vagueness. Yet some authors view the insistence on defining interventions as a dogmatic attitude that stifles research. It is possible that these authors presuppose the existence of "the pure effect" of certain states of nature (e.g., being obese, being black) that does not depend on how that state was reached, which would explain the depiction of these states as treatments or exposures in directed acyclic graphs that represent nonparametric structural equations models.[43] However, for many of us, it is unclear whether such "pure effects" generally exist (after all, different interventions often result in different effect estimates), and therefore it is also unclear whether nonparametric structural questions models are always scientifically meaningful.

A potentially useful function of comparing outcomes under ill-defined interventions is the identification of targets for action.[6] For example, one could argue that comparisons of health outcomes between blacks and whites in the United States measure the causal effect of centuries of oppression, political disenfranchisement, economic warfare, and other aspects of racism. Though an intervention to revert those historical factors is ill-defined (and anyway unfeasible), it might be argued that a white-black comparison identifies racism as a target for intervention. This argument begs the question of "how do we reduce racism?", which can only be answered by sufficiently specifying the feasible interventions that we propose to implement.

Thus a will to compare ill-defined interventions may be a reflection of scientific intellectual curiosity or of a hopeful search for intervention targets. However, when the goal is urgent political change rather than an eventual understanding the world, a comparison of ill-defined interventions is simply inadequate to provide input for decision makers, especially in light of the severe political and economic constraints they often face.

In summary, the potential outcomes framework does not limit the scope of the causal questions; it just makes it transparent the interpretability and reliance on data of our effect estimates. We owe this transparency to those who will ground their decisions on the results of our research.

Acknowledgments

I am indebted to Dana Flanders, Sander Greenland, Jay Kaufman, James Robins, Sharon Schwartz, Ian Shrier, Sonja Swanson, Jan Vandenbroucke, and Tyler Vanderweele for helpful discussions and comments to earlier versions of this commentary.

Funding: NIH R01 AI102634

References

1. Neyman J. On the Application of Probability Theory to Agricultural Experiments: Essay on Principles, Section 9 (1923). Translated in:. Statistical Science. 1990; 5:465–480.
2. Rubin DB. Estimating causal effects of treatments in randomized and nonrandomized studies. Journal of Educational Psychology. 1974; 66(5):688–701.
3. Rubin DB. Bayesian inference for causal effects: The role of randomization. Annals of Statistics. 1978; 6(1):34–58.
4. Robins JM. A new approach to causal inference in mortality studies with a sustained exposure period — Application to the healthy worker survivor effect [published errata appear in Mathl Modelling 1987;14:917-21]. Mathematical Modelling. 1986; 7:1393–1512.
5. Robins JM. Addendum to "A new approach to causal inference in mortality studies with a sustained exposure period — Application to the healthy worker survivor effect" [published errata appear in Computers Math Applic 1989:18;477]. Computers and Mathematics with Applications. 1987; 14:923–945.
6. Schwartz S, Gatto NM, Campbell UB. Causal identification: a charge of epidemiology in danger of marginalization. Annals of Epidemiology. 2016 (in press).
7. Vandenbroucke JP, Broadbent A, Pearce N. Causality and causal inference in epidemiology: the need for a pluralistic approach. Int J Epidemiol. 2016

8. Schwartz S, Prins SJ, Campbell UB, Gatto NM. Is the "well-defined intervention assumption" politically conservative? Soc Sci Med. 2015

9. Daniel RM, De Stavola BL, Vansteelandt S. The formal approach to quantitative causal inference in epidemiology: misguided or misrepresented? International Journal of Epidemiology. 2016 (in press).

10. Robins JM, Weissman M. Counterfactual causation and streetlamps. What is to be done? International Journal of Epidemiology. 2016 (in press).

11. VanderWeele TJ. On causes, causal inference, and potential outcomes. International Journal of Epidemiology. 2016 (in press).

12. Kaufman JS. There is no virtue in vagueness. Annals of Epidemiology. 2016 (in press).

13. Broadbent A. Causation and prediction in epidemiology: A guide to the"Methodological Revolution”. Studies in History and Philosophy of Biological and Biomedical Sciences. 2015; 54:72–80. [PubMed: 26170216]

14. Tchetgen Tchetgen EJ, VanderWeele TJ. On causal inference in the presence of interference. Statistical Methods in Medical Research. 2012; 21(1):55–75. [PubMed: 21068053]

15. Hudgens MG, Halloran ME. Toward causal inference with interference. Journal of the American Statistical Association. 2008; 103(482):832–842. [PubMed: 19081744]

16. Cole SR, Frangakis CE. The consistency statement in causal inference: a definition or an assumption? Epidemiology. 2009; 20(1):3–5. [PubMed: 19234395]

17. VanderWeele TJ. Concerning the consistency assumption in causal inference. Epidemiology. 2009; 20(6):880–883. [PubMed: 19829187]

18. Pearl J. On the consistency rule in causal inference: axiom, definition, assumption, or theorem? Epidemiology. 2010; 21(6):872–875. [PubMed: 20864888]

19. Hernán MA, VanderWeele TJ. Compound treatments and transportability of causal inference. Epidemiology. 2011; 22(3):368–377. [PubMed: 21399502]

20. Robins, JM.; Greenland, S. Comment on "Causal inference without counterfactuals". In: Dawid, AP., editor. Journal of the American Statistical Association. Vol. 95. 2000. p. 431-435.

21. Hernán MA, Robins JM. Using big data to emulate a target trial when a randomized trial is not available. American Journal of Epidemiology. 2016; 183(8):758–764. [PubMed: 26994063]

22. Hernán MA. Counterpoint: Epidemiology to Guide Decision-Making: Moving Away From Practice-Free Research. American Journal of Epidemiology. 2015; 182(10):834–839. [PubMed: 26507306]

23. Hernán MA, Sauer BC, Hernández-Díaz S, Platt R, Shrier I. Specifying a target trial prevents immortal time bias and other self-inflicted injuries in observational analyses. J Clin Epidemiol. 2016

24. Cochran, W. Observational Studies. In: Bancroft, TA., editor. Statistical Papers in Honor of George W. Snedecor. Iowa State University Press; 1972. p. 77-90.

25. Feinstein AR. Clinical biostatistics. XI. Sources of ‘chronology bias’ in cohort statistics. Clinical Pharmacology and Therapeutics. 1971; 12(5):864–879. [PubMed: 5568260]

26. Dawid AP. Causal inference without counterfactuals. Journal of the American Statistical Association. 2000; 95(450):407–424.

27. Greenland, S. Causal Inference as a Prediction Problem. In: Berzuini, C.; Dawid, AP.; Bernardinelli, L., editors. Causal Inference: Statistical Perspectives and Applications. New York: Wiley; 2012. p. 43-58.

28. Hernán MA, Taubman SL. Does obesity shorten life? The importance of well-defined interventions to answer causal questions. Int J Obes (Lond). 2008; 32(Suppl 3):S8–S14.

29. Hernán MA. Invited commentary: Hypothetical interventions to define causal effects: afterthought or prerequisite? American Journal of Epidemiology. 2005; 162:618–620. [PubMed: 16120710]

30. VanderWeele TJ, Hernán MA. Causal inference under multiple versions of treatment. Journal of Causal Inference. 2013; 1(1):1–20. [PubMed: 25379365]

31. Hernán MA. Invited commentary: Agent-based models for causal inference-reweighting data and theory in epidemiology. American Journal of Epidemioly. 2015; 181(2):103–105.

32. Murray EJ, Robins JM, Seage GRI, Freedberg KA, Hernán MA. A comparison of agent-based models and the parametric g-formula for causal inference. American Journal of Epidemiology. 2016 (in press).

33. Marshall BD, Galea S. Formalizing the role of agent-based modeling in causal inference and epidemiology. Am J Epidemiol. 2015; 181(2):92–99. [PubMed: 25480821]

34. Greenland S. Epidemiologic measures and policy formulation: Lessons from potential outcomes (with discussion). Emerging Themes in Epidemiology. 2005; 2:5. [PubMed: 15921514]

35. Glymour C, Glymour MR. Commentary: race and sex are causes. Epidemiology. 2014; 25(4):488–490. [PubMed: 24887161]

36. VanderWeele TJ, Robinson WR. On the causal interpretation of race in regressions adjusting for confounding and mediating variables. Epidemiology. 2014; 25(4):473–484. [PubMed: 24887159]

37. VanderWeele TJ, Robinson WR. Rejoinder: how to reduce racial disparities?: Upon what to intervene? Epidemiology. 2014; 25(4):491–493. [PubMed: 24887162]

38. Glass TA, Goodman SN, Hernán MA, Samet JM. Causal inference in public health. Annual review of public health. 2013; 34:61–75.

39. Hume, D. An Enquiry Concerning Human Understanding. Reprinted and edited 1993. Indianapolis/Cambridge: Hacket; 1748.

40. Dawid, AP. Counterfactuals, hypotheticals and potential responses: a philosophical examination of statistical causality. In: Russo, F.; Williamson, J., editors. Causality and Probability in the Sciences: Texts In Philosophy Series. Vol. 5. London: College Publications; 2007. p. 503-532.

41. Dawid AP. Statistical causality from a decision-theoretic perspective. Annual Review of Statistics and Its Application. 2015; 2:273–303.

42. Lipsitch M, Tchetgen Tchetgen E, Cohen T. Negative controls: a tool for detecting confounding and bias in observational studies. Epidemiology. 2010; 21(3):383–388. [PubMed: 20335814]

43. Pearl, J. Causality : Models, Reasoning, and Inference. 2nd. New York: Cambridge University Press; 2009.