What Would Have Been is Not What Would Be: Counterfactuals of the Past and Potential Outcomes of the Future

Sharon Schwartz
Nicolle M. Gatto
Ulka B. Campbell

Chapter In Press in:

Causality and Psychopathology: Finding the Determinants of Disorders and Their Cures. Edited by Patrick Shrout. New York: American Psychiatric Publishers Inc.

What Would Have Been is Not What Would Be: Counterfactuals of the Past and Potential Outcomes of the Future

Sharon Schwartz
Nicolle M. Gatto
Ulka B. Campbell

For Correspondence:

Sharon Schwartz, Ph.D.
Associate Professor of Clinical Epidemiology
Mailman School of Public Health, room 720b
Columbia University
722 West 168th Street
New York, New York 10032

Nicolle M. Gatto, Ph.D.
Director, TA Group Head, Epidemiology
Safety and Risk Management
Medical Division, Pfizer, Inc.

Ulka B. Campbell, Ph.D.
Associate Director, Epidemiology
Safety and Risk Management
Medical Division, Pfizer, Inc.

What Would Have Been is Not What Would Be: Counterfactuals of the Past and Potential Outcomes of the Future

INTRODUCTION

Epidemiology is often described as the basic science of public health. A mainstay of epidemiologic research is to uncover the causes of disease that can serve as the basis for successful public health interventions (e.g., Institute of Medicine 1988; Millbank Memorial Fund 1976). A major obstacle to attaining this goal is that causes can never be seen but only inferred. For this reason, the inferences drawn from our studies must always be interpreted with caution.

Considerable progress has been made in the methods required for sound causal inference. Much of this progress is rooted in a full and rich articulation of the logic behind randomized controlled trials (Holland 1986). From this work, epidemiologists have a much better understanding of barriers to causal inference in observational studies, such as confounding and selection bias, and our tools and concepts are much more refined.

The models behind this progress are often referred to as counterfactual models. Although researchers may be unfamiliar with them, they are widely (although not universally), accepted in the field. Counterfactual models underlie the methodologies that we all use. Within epidemiology, when people talk about a counterfactual model, they usually mean a potential outcomes model – also known as Rubin's causal model.

As laid out by epidemiologists, the potential outcomes model is rooted in the experimental ideas of Cox and Fisher, for which Neyman provided the first mathematical expression. It was popularized by Rubin who extended it to observational studies, and expanded by Robins to exposures that vary over time (Maldonado and Greenland 2002; Hernan 2004; VanderWeele and Hernan

2006). This rich tradition is responsible for much of the progress we have just noted.

Despite this progress in methods of causal inference, a common charge in the epidemiologic literature is that public health interventions based on the causes we identify in our studies, often fail. Even when they do not fail, the magnitudes of the effects of these interventions are often not what we expected. Levins (1996) provides a particularly gloomy assessment:

> "The promises of understanding and progress have not been kept, and the application of science to human affairs has often done great harm. Public health institutions were caught by surprise by the resurgence of old diseases and the appearance of new ones. … Pesticides increase pests, create new pest problems and contribute to the load of poison in our habitat. Antibiotics create new pathogens resistant to our drugs. (p. 1).

A less pessimistic assessment suggests that although public health interventions may be narrowly successful, they may simultaneously lead to considerable harm. An example is the success of anti-smoking campaigns in reducing lung cancer rates in the United States, while simultaneously increasing smoking and thereby lung cancer rates in less developed countries. This unintended consequence resulted from the redirection of cigarette sales to these countries (e.g., Beaglehole and Bonita 1997).

Ironically, researchers often attribute these public health failures to a narrowness of vision imposed by the same models of causal inference that heralded modern advances in epidemiology and allied social and biological sciences. That is, counterfactual models improve causal inference in our studies, but are held at least partly responsible for the failures of the interventions that follow those studies. Critics not only think that counterfactually based approaches in

epidemiology *do not* provide a sound basis for public health interventions, but that they *can not* (e.g., Shy 1997; McMichael 1999).

While there are many aspects of the potential outcomes model that warrant discussion, here we focus on one narrowly framed question: Is it possible, as the critics contend, that the same models that enhance the validity of our studies can mislead us when we try to intervene on the causes these studies uncover? We think the answer is a qualified "yes".

We will argue that the problem arises not because of some failure of the potential outcomes approach itself, but rather because of unintended consequences of the metaphors and tools implied by the model. We think that the language, analogies, and conceptual frame that enhance the valid estimation of precise causal effects, can encourage unrealistic expectations about the relationship between the causal effects uncovered in our studies and results of interventions based on their removal.

More specifically, we will argue that the unrealistic expectations of the success of interventions arise in the potential outcomes frame because of a premature emphasis on the effects of causal manipulation (understanding what would happen if the exposure were altered) at the expense of two other tasks that must come first in epidemiologic research: (1) causal identification (identifying if an exposure did cause an outcome) and (2) causal explanation (understanding how the exposure caused the outcome). We will describe an alternative approach that specifies all three of these steps - causal identification, followed by causal explanation, and then the effects of causal manipulation. While this alternative approach will not solve the discrepancy between the results of our studies and the results of our interventions, it makes the sources of the discrepancy explicit. The roles of causal identification and causal explanation in causal inference, which we build upon here, have been most fully elaborated by Shadish, Cook and Campbell (2002), heirs to a prominent counterfactual tradition in psychology.

We think that a dialogue between these two counterfactual traditions (i.e., the potential outcomes tradition and the Cook and Campbell tradition) can provide a more realistic assessment of what our studies can accomplish and, perhaps, a platform for a more successful translation of basic research findings into sound public health interventions.

To make these arguments, we will: (1) review the history and principles of the potential outcomes model, (2) describe the limitations of this model as the basis for interventions in the real world, and (3) propose an alternative based on an integration of the potential outcomes model with other counterfactual traditions.

We wish to make clear at the outset that virtually all of the ideas in this paper already appear in the causal inference literature (Morgan and Winship 2007). This paper simply presents the picture we see as we stand on the shoulders of the giants in causal inference.

THE POTENTIAL OUTCOMES MODEL

In the epidemiologic literature, a counterfactual approach is generally equated with a potential outcomes model (e.g., Maldonado and Greenland 2002; Hernan 2004; VanderWeele and Hernan 2006). In describing this model, we will use the term “exposure” to mean a variable we are considering as a possible cause. For ease of discourse, we will use binary exposures and outcomes throughout. Thus, individuals will be either exposed or not, and develop the disease or not.

The concept at the heart of the potential outcomes model is the causal effect of an exposure. A causal effect is defined as the difference between the potential outcomes that would arise for an individual under two different exposure conditions. In considering a disease outcome, each individual has a potential outcome for the disease under each exposure condition. Therefore, when comparing two exposure conditions (exposed or not exposed), there are four

possible pairs of potential outcomes for each individual. An individual can develop the disease under both conditions, only under exposure, only under non-exposure or under neither condition.

Greenland and Robins (1986) used response types as a shorthand to describe these different pairs of potential outcomes. Individuals who would develop the disease under either condition (i.e., whether or not they are exposed) are called "doomed"; those who would develop the disease only if they were exposed are called "causal types"; those who develop the disease only if they are not exposed are called "preventive types"; and, those who do not develop the disease under either exposure condition are called "immune".

Every individual is conceptualized as having a potential outcome under each exposure that is independent of their actual exposure. Their potential outcomes are determined by the myriad of largely unknown genetic, in utero, childhood, adult, social, psychological and biological causes to which they have been exposed, other than the exposure under study.

The effect of the exposure for each individual is the difference between her potential outcome under the two exposure conditions, exposed and not. For example, if an individual's potential outcomes were to develop the disease if exposed but not if unexposed, then the exposure is causal for that individual (i.e., she is a causal type).

Rubin uses the term "treatment" to refer to these types of exposures and describes a causal effect in language that implies an imaginary clinical trial. In Rubin's (1978) terms, "The causal effect of one treatment relative to another for a particular experimental unit is the difference between the result if the unit had been exposed to the first treatment and the result if, instead, the unit had been exposed to the second treatment" (p. 34). One of Rubin's contributions was the

popularization of this definition of a causal effect in an experiment and the extension of the definition to observational studies (Hernan 2004).

For example, the causal effect of smoking one pack of cigarettes a day for a year (i.e., the first treatment) relative to not smoking at all (the second treatment) is the difference between the disease outcome for an individual if he smokes a pack a day for a year compared with the disease outcome in that same individual if he does not smoke at all during this same time interval.

One can think about the average causal effect in a population simply as the average of the causal effects for all of the individuals in the population. It is the difference between the disease experience of the individuals in a particular population if we were to expose them all to smoking a pack a day and the disease experience if we were to prevent them from smoking at all during this same time period.

A useful metaphor for this tradition is that of “magic powder”, where the magic powder can remove an exposure. Imagine we sprinkle an exposure on a population and observe the disease outcome. Imagine then that we use “magic powder” to remove that exposure, and can go back in time to see the outcome in the same population. The problem of causal inference is two-fold - we don’t have magic powder and we can’t go back in time. We can never see the same people at the same time exposed and unexposed. That is, we can never see the same people both smoking a pack of cigarettes a day for a year and, simultaneously, not smoking cigarettes at all for a year.

From a potential outcomes perspective, this problem is conceptualized as a missing data problem. For each individual, at least one of the exposure experiences is missing. In our studies, we provide substitutes for the missing data. Of course our substitutes are never exactly the same as what we want. However, they can provide the correct answer if the potential outcomes of the

substitute are the same as the potential outcomes of the target, the person or population you want information about.

The potential outcomes model is clearly a counterfactual model in the sense that the same person cannot simultaneously experience both exposure and non-exposure. The outcomes of at least one of the exposure conditions must represent a counterfactual, an outcome that would have, but did not, happen.

Rubin (2005), however, objects to the use of the term counterfactual when applied to his model. Counterfactual implies there is a fact (e.g. the outcome *that did occur* in a group of exposed individuals) to which the counterfactual (e.g. the outcome *that would have occurred* had this group of individuals not been exposed) is compared. But for Rubin, there is no fact to begin with. Rather the comparison is between the potential outcomes of two hypothetical exposure conditions, neither of which necessarily reflects an occurrence. The causal effect for Rubin is between two hypotheticals. Thus, in the potential outcomes frame, when epidemiologists use the term “counterfactual”, they mean “hypothetical” (Morgan and Winship 2007). This subtle distinction has important implications as we shall see.

This notion of a causal effect as a comparison between two hypotheticals derives from the rootedness of the potential outcomes frame in experimental traditions. Holland (1986), an early colleague of Rubin and explicator of his work, makes this experimental foundation clear in his summary of the three main tenets of the potential outcomes model.

First, the potential outcomes model studies the effects of causes and not the causes of effects. Thus, the goal is to estimate the average causal effect of an exposure, not to identify the causes of an outcome. For a population, this is the average causal effect, defined as the average difference between two potential outcomes for the same individuals, the potential outcome under exposure A vs.

the potential outcome under exposure B. The desired, but unobservable, true causal effect is the difference in outcome in one population under two hypothetical exposure conditions: if we were to expose the entire population to exposure A versus their outcome if we were to expose them to exposure B. As in an experiment the exposure is treated as if it were in the control of the experimenter; the goal is to estimate the effect that this manipulation would have on the outcome.

Second, the effects of causes are always relative to particular comparisons. One cannot ask questions about the effect of a particular exposure, without specifying the alternative exposure that provides the basis for the comparison. For example, smoking a pack of cigarettes a day can be preventive of lung cancer if the comparison was smoking 4 packs of cigarettes a day, but is clearly causal if the comparison was with smoking 0 packs a day. As in an experiment, the effect is the difference between two hypothetical exposure conditions.

Third, potential outcomes models limit the types of factors that can be defined as causes. In particular, attributes of units (e.g., attributes of people such as gender) are not considered to be causes. This requirement clearly derives from the experimental, interventionist grounding of this model. To be a cause (or at least a cause of interest), the factor must be manipulable. In Holland (1986) and Rubin's terminology, "No causation without manipulation" (p.959). [1]

The focus on the effect of causes, the precise definition of the two comparison groups and the emphasis on manipulability, clearly root the potential outcomes approach in experimental traditions. Strengths of this approach include the clarity of the definition of the causal effect being estimated, and the articulation of the assumptions necessary for this effect to be valid. These assumptions are: (a)

[1] Rubin (1986), in commenting on Holland's 1986 article, is not as strict as Holland in demanding that causes be, by definition, manipulable. Nonetheless, he contends that one cannot calculate the causal effect of a non-manipulable cause and co-authored the "no causation without manipulation" mantra.

the two groups being compared (e.g., the exposed and the unexposed) are exchangeable (i.e., they have the same potential outcomes) and (b) SUTVA (the Stable Unit Treatment Value Assumption) holds. While exchangeability is well understood in epidemiology, the requirements of SUTVA may be less accessible.

Stable Unit Treatment Value Assumption (SUTVA)

A valid estimate of this causal effect requires that the two groups being compared (e.g., the exposed and the unexposed) are exchangeable (i.e., that is there is no confounding), and that SUTVA is reasonable. SUTVA requires that: (a) the effect of a treatment is the same, no matter how an individual came to be treated, and (b) the outcome in an individual is not influenced by the treatment that other individuals receive. In Rubin's (1986) language,

> "SUTVA is simply the a priori assumption that the value of Y [i.e. the outcome] for unit u [e.g., a particular person] when exposed to treatment t [e.g., a particular exposure or risk factor] will be the same no matter what mechanism is used to assign treatment t to unit u and no matter what treatments the other units receive… SUTVA is violated when, for example, there exist unrepresented versions of treatments (Y_{tu} depends on which version of treatment t was received) or interference between units (Y_{tui} depends on whether some other unit u_j received treatment t or t')" (p. 961).

Thus if one were to study the effects of a particular form of psychotherapy, SUTVA would be violated if: (a) there were different therapists with different levels of expertise, or some individuals freely agreed to enter therapy while others agreed only at the behest of a relative and the mode of entry influenced the effectiveness of the treatment (producing unrepresented versions of treatments) (Little and Rubin 2000), or (b) individuals in the treatment group

shared insights they learned in therapy with others in the study (producing interference between units) (Little and Rubin 2000).

The language in which SUTVA is described, the effects of treatment assignment and versions of treatments, is again, indicative of the explicit connection between the potential outcomes model and randomized experiments. To make observational studies as close to experiments as possible, we must ensure that those exposed to the "alternative treatments" (i.e., different exposures) are exchangeable in the sense that the same outcomes would arise if the individuals in the different exposure groups were reversed. In addition, we must ensure that we control all factors that violate SUTVA. We do this by carefully defining exposures or risk factors in terms of narrowly defined treatments that can be manipulated, at least in theory.

To continue our smoking example, one could ask questions about the average causal effect of smoking a pack of cigarettes a day for a year (treatment A) compared with never having smoked at all (treatment B) in an observational study. Since we cannot observe the same people simultaneously under two different treatments, we compare the disease experience of two groups of people: one with "treatment A", the exposure of interest, and one with "treatment B", the substitute for the potential outcomes of the same group under the second treatment option. In order for the substitution to yield an accurate effect estimate (i.e. for exchangeability to hold), we must ensure that the smokers and non-smokers are as similar as possible on all causes of the outcome (other than smoking). This can be accomplished by random assignment in an RCT. To meet SUTVA assumptions, we have to: (a) be vigilant to define our exposure precisely so there is only one version of each treatment and be certain that how individuals entered the smoking and non-smoking group did not influence their outcome, and (b) ensure the smoking habits of some individuals in our study did not influence the smoking habits of other individuals.

Barring other methodological problems, it would be assumed that if we did the intervention in real life, that is if we prevented people from smoking a pack of cigarettes a day for a year, the average causal effect estimated from our study would approximate this intervention effect. The potential outcomes model is an attempt to predict the average causal effect that would arise (or be prevented) from a particular manipulation under SUTVA. It is self-consciously interventionist.

Indeed, causal questions are framed in terms of intervention consequences. To ensure the validity of the causal effects uncovered in epidemiologic studies, researchers are encouraged to frame the causal question in these terms. As a prototypical example, Glymour (2007), in a cogent methodological critique of a paper examining the effect of childhood socio-economic position on adult health, re-stated the goal of the study in potential outcome terms. "The primary causal question of interest is how adult health would differ if we intervened to change childhood socio-economic position" (p. 566).

It is critical to note that even when we do not explicitly begin with this type of model, the interventionist focus of the potential outcomes frame implicitly influences our thinking through its influence on our methods. For example, this notion is embodied in our understanding of the attributable risk as the proportion of disease that would be prevented if we were to remove this exposure (Last 2001). More generally, authors often end study reports with a statement about the implications of their findings for intervention or policy that reflect this way of thinking.

Limitations of the Potential Outcomes Model for Interventions in the Real World

To ensure the internal validity of our inferences, we isolate the effects of our causes from the context in which they act. We do this by narrowly defining our

treatments, creating exchangeability between treated and untreated people, and considering social norms and the physical environment as part of a stable background in which causes act. In order for the causal effect of an exposure in a study to translate to the effect of its intervention, all of the controls and conditions imposed in the study must hold in the intervention and the targeted population (e.g. treatment definition, follow-up timeframe, distribution of other causes, etc.).

The problem is that in most cases, interventions in the real world cannot replicate the conditions that gave rise to the average causal effect in a study. It is important to note that that this is true for randomized controlled trials as well as observational studies. It is true for classic risk factors, as well as for exposures in life course and social epidemiology. The artificial world that we appropriately create to identify causal effects - a narrow swath of temporal, geographic and social reality in which exchangeability exists and SUTVA is not violated - captures a vital but limited part of the world in which we are interested. Thus, while the approach we use in studies aids in the valid estimation of a causal effect for the past, it provides a poor indicator of a causal effect for the future. For these reasons, the causal effects of our interventions in the real world are unlikely to be the same as the causal effects of our studies.

This problem is well recognized in the literature on randomized controlled trials in terms of the difference between efficacy and effectiveness and in the epidemiology literature as the difference between internal and external validity. However, this recognition is rarely reflected in research practice. We suspect this problem may be better understood by deeper examination of the causes of the discrepancy between the effects observed in studies and the effects of interventions. We group these causes into three interrelated categories: direct violations of SUTVA, unintended consequences of our interventions, and context dependencies.

Direct Violations of SUTVA

Stable treatment effect : In order to identify a causal effect, a necessary SUTVA assumption is that there is only one version of the treatment. To meet this assumption, we need to define the exposures in our studies in an explicit and narrow way. For example, we would ask about the effects of a particular form of psychotherapy (e.g., interpersonal psychotherapy conducted by expert clinicians) rather than about psychotherapy in general. This is because the specific types of therapy encompassed within the broad category of "psychotherapy" are likely to have different effects on the outcome.

While this is necessary for the estimation of precise causal effects in our studies, it is not likely to reflect the meaning of the exposure or treatment in the real world. The removal of causes or the provision of treatments, no matter how well defined, is never surgical. Unlike the removal of causes by the "magic powder" in our thought experiments, interventions are often crude and messy. Public health interventions are inherently broad. Even in a clinical context, treatment protocols are never followed precisely in real-world practice.

In public health interventions, there are also different ways of getting into "treatment" and these may well have different effects on the outcome. For instance, the effect of an intervention offering a service may be very different for those who use it only after it has become popular (the late adopters). Early adopters of a low fat diet, for example, may increase their intake of fruits and vegetables to compensate for the caloric change. Late adopters may substitute low fat cookies instead. Low fat diet was adopted by both types of people, but the effect on an outcome (e.g., weight loss) would likely differ. There are always different versions of treatments, and the mechanisms through which individuals obtain the treatments will frequently impact the effect of the treatments on the outcome.

Interference between units: When considered in real world applications over a long enough time frame, there will always be “interference between units”. Because people live in social contexts, their behavior influences norms and social expectations. Behavior is contagious. This can work in positive ways, increasing the effectiveness of an intervention, or lead to negative unintended consequences. An example of the former would be when the entrance of a friend into a weight loss program encourages another friend to lose weight (Christakis and Fowler 2007). Thus the outcome for one individual is contingent on the exposure of another individual. Similarly, changes in individual eating behaviors spread. This influences not only individuals’ behavior but, eventually, the products that stores carry, the price of food, and the political clout of like minded individuals. It changes the threshold for the adoption of healthy eating habits. There is an effect not only of the weight loss program itself, but of the proportion of people enrolled in weight loss programs within the population.

Within the time frame of our studies, the extant norms caused by interactions among individuals and the effect of the proportion of exposure in the population, are captured as part of the background within which causes act, are held constant, and are invisible. To identify the true effects these causes had, this approach is reasonable and necessary. The causes worked during that time frame within that normative context. But in a public health intervention, these norms change over time due to the intervention. This problem is well recognized in infectious disease studies where the contagion occurs in a rapid time frame, making non-interference untenable even in the context of a short term study. But it is hard to imagine any behavior which is not contagious over long enough time frames. The fact is that the causal background we must hold constant to estimate a causal effect is influenced by our interventions.

Unintended Consequences of Interventions

Unintended consequences of interventions are consequences of exposure removal not represented as part of the causal effect of the exposure on the outcome under study. The causes of these unintended consequences include natural confounding and narrowly defined outcomes.

Natural confounding: Recall that the estimation of the true causal effect requires exchangeability of potential outcomes between the exposed and unexposed groups in our studies. Exchangeability is necessary to isolate the causal effect of interest. For example, in examining the effects of alcohol abuse on vehicular fatalities, we may control for the use of illicit drugs. We do so because those who abuse alcohol may be more likely to also abuse other drugs that are related to vehicular fatalities.

If the association between alcohol abuse and illicit drug use is a form of "natural confounding", that is the association between alcohol and drug use arises in naturally occurring populations and is not an artifact of selection into the study, then this association is likely to have important influences in a real world intervention. That is, the way in which individuals came to be exposed, may influence the effect of the intervention, in violation of SUTVA.

For example, when two activities derive from a similar underlying factor (social, psychological or biologic), the removal of one may influence the presence of the other over time; it may activate a feedback loop. Thus the causal effect of alcohol abuse on car accidents may overestimate the effect of the removal of alcohol abuse from a population if the intervention on alcohol use inadvertently increases marijuana use.

As this example illustrates, an intervention may influence not only the exposure of interest, but also other causes of the outcome that are linked with the

exposure in the real world. In our studies, we purposely break this link. We overcome the problem of the violation of SUTVA by imposing narrow limits on time and place so that SUTVA holds in the study. We control these variables, precisely because they are also causes of the outcome under study. But in the real world, their influence may make the interventions less effective than our effect estimates suggest. The control in the study was not incorrect, as it was necessary to isolate the true effect that alcohol use did have on car accidents among these individuals given the extant conditions of the study. But outside the context of the study, removal of the exposure of interest had unintended consequences over time through its link with other causes of the outcome.

Narrowly defined outcomes: Although we may frame our studies as identifying the "effects of causes", they only identify the effects of causes on the specific outcomes we examine in our studies. In the real world, causes are likely to have many effects. Likewise, our interventions have effects on many outcomes, not only those we intend. Unless we consider the full range of outcomes, our interventions may be narrowly successful, but broadly harmful. For example, successful treatments for AIDS have decreased the death rate, but have also led people to re-conceptualize AIDS from a lethal illness to a manageable chronic disease. This norm change can lead to a concomitant rise in risk taking behaviors and an increase in disease incidence. More optimistically, our interventions may have beneficial effects that are greater than we assume if we consider unintended positive effects. For example, an intervention designed to increase high school graduation rates may also reduce alcoholism among teens.

Context Dependency

Most fundamentally, all causal effects are context dependent and therefore all effects are local effects. It is unlikely that a public health intervention will be applied only in the exact population in which the causal effects were studied. Public health interventions often apply to people who do not volunteer for them,

to a broader swath of the social fabric and over a different historical time frame. Therefore, even if our effect estimates were perfectly valid, we would expect effects to vary between our studies and our interventions. For example, psychiatric drugs are often tested on individuals who meet strict DSM criteria, do not have comorbidities and are placebo non-responders. Once the drugs are marketed, however, they are used to treat individuals who represent a much wider population. It is unlikely that the effects of the drugs will be similar in its real world usage as in the studies.

For all these reasons, it seems unlikely that the causal effect of any intervention will reflect the causal effect found in our studies. These problems are well known and much discussed in the social science literature (e.g., Merton 1936; 1968; Lieberson 1985) and the epidemiologic literature (e.g., Greenland 2005).

Nonetheless, when carrying out studies, epidemiologists often talk about trying to identify "the true causal effect of an exposure", as if this was a quantification that has some inherent meaning. An attributable risk is interpreted as if this provided a quantification of the effect of the elimination of the exposure under study. Policy implications of etiologic work are discussed as if they flowed directly from our results. We think that this is an overly optimistic assessment of what our studies can show. We think that as a field, we tend to estimate the effect exposures had in the past and assume that this will be the effect in the future. We do this by treating the counterfactual of the past as equivalent to the potential outcome of the future.

AN ALTERNATIVE COUNTERFACTUAL FRAMEWORK (AN INTEGRATED COUNTERFACTUAL APPROACH)

An alternative framework, which we will refer to as an Integrated Counterfactual Approach (ICA) distinguishes three sequential tasks in the relationship between

etiologic studies and public health interventions, the first two of which are not explicit goals in a potential outcomes frame: 1) causal identification, 2) causal explanation, and 3) the effects of causal manipulation.

Step 1: Causal Identification

In line with the Cook and Campbell tradition (Shadish, Cook and Campbell 2002), this alternative causal approach uses the insights and methods of potential outcomes models, but reframes the question that these models address as the identification of a cause rather than the result of a manipulation. Whereas the potential outcomes model is rooted in experiments, the Integrated Counterfactual Approach (ICA) is rooted in philosophic discussions of counterfactual definitions of a cause, particularly the work of Mackie (1965; 1974). It begins with Mackie's definition of a cause rather than a definition of a causal effect.

For Mackie, X is a cause of Y if, within a causal field, with all held constant, Y would not have occurred if X had not, at least not when and how it did. Mackie's formulation begins with a particular outcome and attempts to identify some of the factors that caused it. Thus the causal contrast for Mackie is between what actually happened and what would have happened had everything remained the same except that one of the exposures was absent. The contrast represents the difference between a fact and a counterfactual, rather than two potential outcomes.

Thus, for Mackie, something is a cause if the outcome under exposure is different from what the outcome would have been under non-exposure. By beginning with actual occurrences, Mackie gives prominence to the contingency of all causal identification. This approach explicitly recognizes that causes are always identified within a causal field of interest, where certain factors are assumed to be part of the background in which causes act, rather than factors

available for consideration as causes. The decision to assign factors to the background may differ among researchers and time periods. Thus there is a subjective element in deciding which, among the myriad of possible exposures, a factor is hypothesized to be a cause of interest.

Rothman and Greenland (1998) provide a definition of a cause in the context of health that is consistent with Mackie's view.

> "We can define a cause of a specific disease event as an antecedent event, condition, or characteristic that was necessary for the occurrence for the disease at the moment it occurred, given that other conditions are fixed." (p. 8).

As applied to a health context, both Mackie and Rothman and Greenland begin with the notion that for most diseases, an individual can develop a disease from one of many possible causes, each of which consists of several components working together. In this model, although none of the components in any given constellation can cause disease by itself, each makes a non-redundant and necessary contribution to complete a causal mechanism. A constellation of components that is minimally sufficient to cause disease is termed a sufficient cause. Mackie referred to these component causes as INUS causes (Insufficient but Necessary components of Unnecessary but Sufficient causes). Rothman's (1976) sufficient causes are typically depicted as "causal pies".

As an example, let's assume that the disease of interest is schizophrenia. There may be three sufficient causes of this disease (see Figure 1). An individual can develop schizophrenia from a genetic variant, a traumatic event and poor nutrition; or from stressful life events, childhood neglect and exposure to an environmental toxin; or from prenatal viral exposures, childhood viral exposure and a vitamin deficiency. We have added components U_1, U_2, and U_3 to the sufficient causes to represent additional unknown factors. Each individual

develops schizophrenia from one of these sufficient causes; in no instance does the disease occur from any one factor - rather it occurs due to several factors working in tandem.

The ICA and potential outcomes model are quite consistent in many critical ways in this first step. Indeed, the potential outcomes model provides a logical, formal, statistical framework applicable to causal inference within the context of the ICA. Regardless of whether we intend to identify a cause or estimate a causal effect, the same isolation of the cause is required. Most essentially, this means that comparison groups must be exchangeable. However, each framework is intended to answer different questions (see Table 1).

From a potential outcomes perspective, the goal is to estimate the causal effect of the exposure. From an ICA perspective, the goal is to identify whether an exposure was a cause. This distinction between the goals of identifying the effects of causes and the causes of effects is critical and has many consequences.

First, identifying the effects of causes is future-oriented. We start with a cause and estimate its effect. The causal contrast is between the potential disease experiences of a group of individuals under two exposure conditions. Identifying the causes of effects, in contrast, implies that the identification is about what happened in the past. The causal contrast is between what did happen to a group of individuals under the condition of exposure, something explicitly grounded in and limited by a particular socio-historical reality, and what would have happened had all conditions remained constant except that the exposure was absent. This approach identifies factors that actually were causes of the outcome. Whether or not they will be causes of the outcome depends on the constellation of the other factors held constant at that particular socio-historical moment. The effect of this cause in the future is explicitly considered a separate question.

Second, when we consider a potential outcomes model, the causal effect of interest is most often the causal effect for the entire population. That is, we conceptualize the causal contrast as the entire study population under two different treatments. We create exchangeability by mimicking random assignment. Neither exposure condition is "fact" or "counterfactual". Rather, both treatment conditions are substitutes for the experience of the entire population under that treatment.

In contrast, Mackie's perspective implies that the counterfactual of interest is a counterfactual for the exposed. We take as a given, what actually happened to people with the putative causal factor and imagine a counterfactual in reference to them. We create exchangeability by mimicking the predispositions of the exposed. This puts a different spin on the issue of confounding and non-exchangeability.[2] The factors that differentiate exposed and unexposed people are more easily seen as grounded in characteristics of truly exposed people and their settings. It makes explicit that the causal effect for people who are actually exposed may not be the same as the effect that cause would have on other individuals. Thus this type of confounding is seen not as a study artifact but as a form of true differences between exposed and unexposed people that can be and must be adjusted for in our study, but must also be considered as an active element in any real life intervention.

Third, the focus on estimating the effects of causes in the potential outcomes model, leads to the requirement of manipulability; any factor which is not manipulable is not fodder for casual inference. From an ICA perspective, any factor can be a cause (Shadish, Cook and Campbell 2002). To qualify, it has to be a factor that, were it absent, with all else the same, this outcome, within this context would not have occurred. Even characteristics of individuals, such as

[2] Technically, when the effect for the entire population is of interest, full exchangeability is required. When the effect for the exposed is of interest, only partial exchangeability is required (Greenland and Robins 1986).

gender, are grist for a counterfactual thought experiment. The world is fixed as it is in this context, say with a fairly rigid set of social expectations depending on identified sex at birth. We can ask a question about what an individual's life would have been like had they been born male, rather than female, given this social context.

Fourth, this perspective brings the issue of context dependence front and center. As Rothman's (1976) and Mackie's (1965) models make explicit, shifts in the component causes and their distributions, variations in the field of interest and the socio-historical context, change the impact of the cause and indeed, determine whether or not the factor is a cause in this circumstance. Thus the impact of a cause is explicitly recognized as context dependent; the size of an effect is not universal. A factor can be a cause for some individuals in some contexts, but not in others. Thus the goal is the "identification of causes in the universe", rather than the estimation of universal causal effects. By "causes in the universe", we mean factors which at some moment in time have caused the outcome of interest and could theoretically (if all else were equal), happen again.

Step 2: Causal Explanation

The focus on the causes of effects facilitates an important distinction that emerges from the Cook and Campbell (1979) tradition - that between causal identification and causal explanation. From their perspective, in the first step, we identify whether the exposure of interest did cause the outcome in some people in our study. We label this "causal identification"[3]. If we want to understand the effect altering a cause in the future, an additional step of causal explanation is required. Causal explanation comprises two components, construct validity, an understanding of the "active ingredients" of the exposure

[3] Shadish, Cook and Campbell (2005) call this step, causal description. We think the term causal identification is a better fit for our purposes.

and how they work, and external validity, an identification of the characteristics of persons, places and settings that facilitate its effect on the outcome.

Construct validity

In causal identification, we examine the causal effects of our variables as measured. In causal explanation, we ask what it is about these variables that caused the outcome. Through mediational analyses, we examine both the active ingredients of the exposure (i.e., what aspects of the exposure are causal) and explore the pathways through which the exposure affects the outcome. Mediational analyses explicitly explore the potential SUTVA violation inherent in different versions of treatments. Exploration of pathways can lead to a more parsimonious explanation for findings across different exposure measures. Based on the active ingredients of exposure (and their resultant pathways), we can test not only the specific exposure-disease relationship, but also a more integrative theory regarding the underlying "general causal mechanisms" (Judd and Kenny 1981). This theory allows us to make statements about an observed association that are less bounded by the specific circumstances of a given study, and generalize based on deep similarities (Judd and Kenny 1981; Shadish, Cook et al. 2002). This generalization has two practical benefits. First, knowledge of mechanisms enhances our ability to compare study results across exposures, and thus integrate present knowledge. Second, such an analysis can help to identify previously unknown exposures or treatments, due to the fact that they capture the same active ingredient (or work through the same mechanism) as the exposure or treatment under study) (Hafeman 2008).

Let's continue the gender example. First we try and test the hypothesis that female gender was a cause of depression for some people in our sample. By this we mean that there are some people who got depressed as women who would not have been depressed had they not been women (i.e. if they were male –or some other gender). Of course causal inference is tentative as always. But

let's assume that at this first step we identified something that is not just an association, we took care to rule out all non-causal alternative explanations to the best of our ability.

Once that step is accomplished, we may ask how female gender causes disease. Gender is a multifaceted construct with many different aspects, genetic, hormonal, psychological and social. Once we know that gender has a causal effect, probing the construct helps us to identify what it is about female gender that causes depression. This may help verify gender's causality in depression, and help us to identify other exposures that do the same thing as gender (that is other constructs that have the same active ingredient). For example, some have suggested that the powerlessness of women's social roles is an active ingredient in female gender as a cause of depression. This would suggest that other social roles related to powerlessness, such as low socio-economic position, might also be causally related to depression. Probing the construct of the outcome plays a similar role in causal explanation. It helps to identify the specific aspects of the outcome that are influenced by the exposure and helps to refine the definition of the outcome.

External Validity

The other aspect of causal explanation requires an examination of the conditions under which exposures act (Shadish, Cook and Campbell 2002). The context dependency of causal effects is therefore made explicit. Causal inference is strengthened through the theoretical consideration and testing of effect variation. From the perspective of the ICA, consistency of effects across settings, people and time periods is not the expectation. Rather variation is expected and requires examination.

When we identify causes in our studies, we make decisions about the presumptive world that we hold constant, considering everything as it was when

the exposure arose. Thus the social effects and norms that may have been consequences of the exposure are frozen in the context. But when we intervene on our causes, we must consider the new context. This aspect of causal explanation, the specification of the conditions under which exposures will and will not cause disease, is considered the separate task of external validity in the Cook and Campbell scheme.

Step 3: Causal Manipulation

While this separation of causal identification and causal explanation has the benefit of placing contingency and context dependence center stage it does not resolve the discrepancy between the effects observed in studies and the effects of interventions. It does not provide the tools necessary to uncover the feedback loops and unintended consequences of our interventions. It does not fully address the violation of the SUTVA assumption of no interference between units. Even causal explanation is conducted within established methods of isolation, reductionism and linearity. Prediction of the effects of causal manipulation may require a different approach, one rooted in complexity theories and systems analysis, as the critics contend (e.g., McMichael 1999; Levins 1997: Krieger 1994).

To understand an intervention, the complexity of the system and feedbacks depends, of course, on the question at hand. The critical issue, as Levins (1996) notes, is the ability to decide when simplification is constructive and when it is an obfuscation.

The implementation of systems approaches within epidemiology requires considerable methodological and conceptual development, but may be a required third step to link etiologic research to policy.

The integrated causal approach does not provide a solution to the discrepancy between the results of etiologic studies and the results of public health interventions. But it does provide a way of thinking in which causal identification is explicitly conceptualized as a first step rather than a last step for public health intervention. It is a road map to a proposed peace treaty in the epidemiology wars between the counterfactual and dynamic model camps. It suggests that the models are useful for different types of questions. Counterfactual approaches, under SUTVA, are essential for identifying causes of the past. Dynamic models allowing for violations of SUTVA are required to understand potential outcomes of the future.

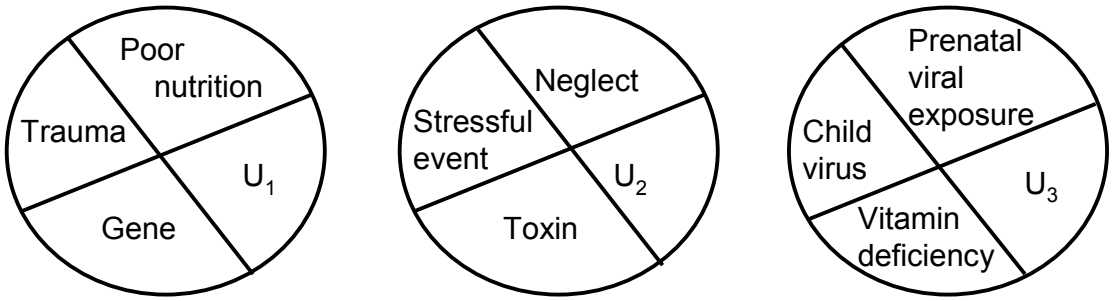
SUMMARY

The rigor of causal inference, brought to light in the development of the potential outcomes model, is essential as the basis for any intervention. Rigor is demanded, because interventions developed around non-causal associations are doomed to failure. However, reifying the results of our studies by treating causes as potential interventions is also problematic.

We suspect that public health will benefit from interventions identified using an approach that integrates the potential outcomes tradition of Rubins and Robins in statistics and epidemiology with the counterfactual tradition of Shadish, Cook, and Campbell in psychology. This integrated approach clarifies that the identification of causes facilitated by isolation is only a first step in policy formation. A second step, causal explanation, aids in the generalizability of our findings. Here, however, instead of replication of our study in different contexts, we generalize on the basis of the deep similarities uncovered through casual explanations. The steps of identification and explanation may require a third step of prediction to understand intervention effects. The causes that we identify, together with their mediators and effect modifiers, may be considered nodes in

more complex analyses that allow for the consideration of feedback loops and the unintended consequences that are inherent in any policy application. The methods for this final step have not yet been fully developed. The conceptual separation of these three questions, grounded in a distinction between counterfactuals of the past and potential outcomes of the future may prepare the ground for such innovations. For as Kierkegaard (1943; cited in Hannaly 1996) noted, “life is to be understood backwards, but it is lived forwards”. At a minimum, we hope that a more modest assessment of what current epidemiology methods can provide will help stem cynicism that inevitably arises when we promise more than we can possibly deliver.

Figure 1: Potential causes of schizophrenia depicted as causal pies



Sufficient Cause 1 | **Sufficient Cause 2** | **Sufficient Cause 3**

Adapted from Rothman and Greenland, 1998

Table 1: Differences between the Potential Outcomes Model and an Integrated Counterfactual Approach

	Potential Outcomes Model	**Integrated Counterfactual Approach**
Goal • Salient differences	Estimation of true causal effect • Estimate • Quantitative • Effects of causes	Identification of true causes • Identify • Qualitative • Causes of effects
Means • Salient differences	Compare two potential outcomes • Entire population under two exposures • Manipulable causes • SUTVA • Mimic random assignment	Compare a fact with a counterfactual • Exposed under two exposure conditions • Any factor • Construct validity • Mimic assignment of exposed
Interpretation • Salient differences	Potential outcome of the future • Expect consistency	Causal effect of the past • Expect inconsistency
Roots	Experiments / Cox, Neyman, Fisher	Counterfactual philosophy / Mackie, Cook & Campbell

Citations

Beaglehole R and Bonita R (1997). *Public Health at the Crossroads: Achievements and Prospects.* New York: Cambridge University Press.

Cassel J and Lebowitz MD. 1976. Causality in the environment and health: the utility of the multiplex variables. *Perspectives in Biological Medicine* 19:338-343.

Christakis NA and Folwer JH. (2007). The spread of obesity in a large social network over 32 years. *The New England Journal of Medicine*. 357:370-379.

Cook TD and Campbell DT. (1979) *Quasi-experimentation: Design and Analysis Issues for Field Settings.* Chicago: Rand McNally.

Glymour MM. (2007) Selected samples and nebulous measures: some methodological difficulties in life-course epidemiology. *International Journal of Epidemiology*. 36:566-568.

Greenland S, Robins JM. (1986) Identifiability, exchangeability and epidemiological confounding. *International Journal of Epidemiology*. 15:413-419.

Hafeman D. (2008) *Opening the Black Box: a Re-Assessment of Mediation from a Counterfactual Perspective*. Ph.D. dissertation, Columbia University.

Hannay A. (1996) *Soren Kierkegaard S. (1843) : Papers and Journals*. London: Penguin Books.

Hernan MA. (2004). A definition of causal effect for epidemiological research. *Journal of Epidemiology and Community Health* 58:265-271.

Holland PW (1986) Statistics and Causal Inference. *Journal of the American Statistical Association.* 81:945-960.

Institute of Medicine (1988). Committee for the study of the future of public health, division of health care services. *The Future of Public Health*. National Academy Press: Washington, D.C.

Judd CM and Kenny DA. (1981) Process analysis: estimating mediation in treatment evaluations. *Evaluation Review* 5: 602-619.

Krieger N. (1994) Epidemiology and the web of causation: has anyone seen the spider? *Social Science and Medicine* 39:887-903.

Last JM. (2001). *A Dictionary of Epidemiology*. New York: Oxford University Press.

Levins R (1997) When science fails us. *Forests, Trees and People Newsletter* 32/33: 1-18.

Levins (1996). Ten propositions on science and anti-science. *Social Text* 46/47: 101-111.

Lieberson S. (1985). *Making it Count: The Improvement of Social Research and Theory.* Berkeley: University of California Press.

Little RJ, Rubin DB. (2000) Causal effects in clinical and epidemiological studies via potential outcomes: concepts and analytical approaches. Annual Review of Public Health 21:121-145.

Mackie JL (1965) Causes and conditions. *American Philosophical Quarterly* 4:245-264.

Mackie JL. (1974) *Cement of the Universe: A Study of Causation*. Oxford: Oxford University Press.

McMichael AJ (1999) Prisoners of the Proximate: Loosening the Constraints on Epidemiology in an Age of Change. *American Journal of Epidemiology* 149:887-897.

Merton RK. (1936). The unanticipated consequences of purposive social action. *American Sociological Review* 1:894-904

Merton, RK. (1968) *Social Theory and Social Structure*. New York: The Free Press.

Milbank Memorial Fund Commission (1976). H*igher Education for Public Health: A Report of the Milbank Memorial Fund Commission.* New York: Prodist.

Morgan SL and Winship C. (2007). *Counterfactuals and Causal Inference: Methods and Principles for Social Research.* Cambridge: Cambridge University Press.

Rothman KJ. 1976 Causes. *American Journal of Epidemiology* 104: 587-592.

Rubin DB (1978) Bayesian inference for causal effects: The role of randomization. *The Annals of Statistics* 6: 34-58.

Rubin DB (1986) Statistics and causal inference comment: which ifs have causal answers. *Journal of the American Statistical Association* 81: 961-962.

Rubin DB (2005). Causal inference using potential outcomes: design, modeling decisions. *Journal of the American Statistical Association* 100:322-331.

Sander Greenland (2005) Epidemiologic measures and policy formulation: lessons from potential outcomes. *Emerging Themes in Epidemiology* 2:1-7

Shadish WR, Cook TD, Campbell DT. (2002) *Experimental and Quasi-experimental Designs for Generalized Causal Inference*. Boston: Houghton Mifflin Company.

Shy CM (1997) The failure of academic epidemiology: witness for the prosecution. *American Journal of Epidemiology* 145:479-484.

VanderWeele TJ and Hernan MA. (2006) From counterfactuals to sufficient component causes and vice versa. *European Journal of Epidemiology* 21:855-858.