**Causal Attribution, Counterfactuals, and Disease Interventions**

**James Woodward**

**HPS Pittsburgh**

Forthcoming in *Measuring the Global Burden of Disease: Philosophical Dimensions* (Oxford University Press)

**1. Introduction**

This paper explores a number of interrelated issues that affect assessment of the global burden of disease: the causal attribution of disability and death to particular diseases, the relation of such causal attribution information to predicting the effects of interventions to remove or reduce the incidence of these diseases (disease interventions), and the use of counterfactuals in epidemiological causal reasoning. Section 2 begins by discussing causal attribution data and their connection to predicting the effects of disease interventions on death. Section 3 argues that we can reliably predict the effects of these disease interventions from causal attribution data alone only when strong additional “independence” assumptions are satisfied. It also argues that in many realistic circumstances these assumptions are unlikely to hold, and that when they do not, additional information besides that provided by causal attribution data is needed in order to predict the effects of disease interventions. Among other things, one needs to explicitly model the causal relationships among different diseases or causes of death. This in turn requires frameworks, e.g. structural equations and directed graphs, that explicitly incorporate counterfactual information about what would happen if one were to intervene in various ways. Sections 4 -6 describe such frameworks and illustrate their application to some simple examples. Section 7 further explores the relation between causal attribution information and claims about the outcomes of interventions. Section 8 discusses the sorts of variables that can figure in causal claims in epidemiology. Section 9 addresses the interpretation of counterfactuals. Section 10 concludes with discussion of the relevance of the epidemiological notions of excess and etiological fractions to these matters.

 **2. Attribution of Deaths to Causes**

As Mathers et al. (2002) explain, an important goal of summary measures of public health (SMPH) and the Global Burden of Disease (GBD) project more generally is understanding the causal determinants of death and disability. One goal of the GBD project is to use such causal information to predict the effects of various possible public health interventions on such causes and in this way to prioritize programs and assess them in cost benefit terms. Standard measures of the impact of various diseases on health and disability rely on such intervention-based predictions. They concern, for example, the expected number of years of life that might be gained in a population if a certain disease were eliminated.

As Mathers et al. (2002) also explain, the causal information on which researchers working within the GBD project attempt to base such results of intervention-based predictions often comes from what they call “categorical attribution” of causes of death. Here “an event such as death is attributed to a single cause, according to a set of rules” (Mathers, 2002 p. 274)—for example, the rules specified in the International Classification of Diseases, which specify when a death is to be attributed to, e.g., a heart attack or to lung cancer. A central feature of such rules is that they specify a set of causes that are mutually exclusive and exhaustive: each death is attributed to one and only one cause. (p.274). Categorical attribution information about cause of death can be obtained from death certificates and from so-called verbal autopsies. The gold standard is presumably something like an actual autopsy.

Mather et al. (2002) contrast such causal attribution information with what they call “counterfactual analysis” in which “the contribution of a diseases, injury or risk factor” is estimated by comparing the current future levels of a summary measure [e.g. deaths or years of life lost or disability adjusted years lost] with the levels that would be expected under some alternative, hypothetical or counterfactual scenario”. Estimation of such measures generally involves the use of causal modeling techniques applied to population level information about the incidences of various causes of death and disability in a population, and the observed patterns of association among these. This information then is used to construct a model describing the causal relationships among these quantities—both among the causes of death and disability and their connection to these outcomes.

Although Mather el al. (2002) contrast causal attribution with “counterfactual analysis” it is worth emphasizing that the sorts of questions that arise in connection with predicting the results of disease interventions (whether these are based on causal attribution information or something else) inevitably concern the assessment of counterfactuals —they have to do, for example, with what would happen if, contrary to what is presently the case, a disease intervention program with certain characteristics were to be introduced. So to the extent that predicting the outcomes of such programs is the goal, some form of counterfactual analysis seems unavoidable, with the only issue being whether and to what extent the use of causal attribution information can contribute to assessing such counterfactual claims[[1]](#footnote-1).

 This motivates the following natural and important question: if one’s goal is ultimately to reliably predict the result of disease interventions or to answer counterfactual questions about what would happen if such interventions were to be undertaken, what can causal attribution information tell us that is relevant to this goal? Put slightly differently, in what circumstances can we infer from information concerning deaths (or other adverse outcomes) attributed to some cause *C* to conclusions about what would happen to those outcomes if we were to eliminate or at least reduce the incidence of *C*?

To explore this question, it will be helpful to introduce some terminology and distinctions. It is standard in the philosophical literature to contrast *type* and *token* causal claims (or, as the latter are sometimes —and I believe more appropriately—called *singular causal* or *actual cause* claims). The latter are claims about the causation of particular individual events or outcomes by particular causes –e.g., that Jones’ lung cancer caused his death or that Jones’ smoking caused his lung cancer. Causal attribution claims about causes of particular deaths appear to be (one particular kind of) token or actual cause claim. By contrast, type causal claims can be understood as (in some way) claims about the types of effects produced by types of causes (in some cases, claims concerning causal relationships in a particular population or set of populations; in other cases, claims about the tendency or capacity of types of causes to produce types of effects but with no reference to particular populations), with no implications about the causation of any particular outcome such as the death of a particular individual.

Examples of type causal claims are “Smoking causes lung cancer” and “lung cancer causes death” (no particular populations specified) or “Smoking caused an increase in lung cancer in the U.S population in the middle decades 20th century” or “Such and such a percentage of deaths in the U.S population in 2010 were caused by smoking” (references to a particular population). As these examples illustrate, type causal claims can be more or less quantitatively precise. Adopting a somewhat expansive conception of the notion of a type causal claim, I think of many of the causal claims made in epidemiology, including claims about relative risks of death or disease caused by exposure or claims about excess or attributable fractions, when these are interpreted causally and not as mere measures of association*[[2]](#footnote-2)*, are type causal claims, in the sense that they are not claims about the causation of outcomes in particular identified individuals. Claims about the effects of disease intervention programs are also naturally understood as type-like claims, typically concerning particular populations.

 Type causal claims can be supported by evidence from many different sources but typically in epidemiology this evidence includes statistical information of various sorts concerning patterns of correlation among sets of variables[[3]](#footnote-3). Conversely, type level claims are usually understood as having implications regarding correlations or statistical associations—e.g. for the existence of correlations between smoking and lung cancer in various populations.

As Ned Hall (in his oral presentation at this conference) and others (myself included—see Woodward, 2018) have observed, the terminology of type and token causation, although well-entrenched, is problematic in a number of respects. For one thing, in a so-called type claim like “smoking causes lung cancer”, the relevant units are still particular individuals – it is particular individuals with individual histories of smoking or not who get or fail to get lung cancer. There is not some mysterious causal relationship between the *types* “smoking” and “lung cancer” over and above this, as the terminology of type-causation might seem to suggest. In other words, whether we make type or token causal claims about the connection between smoking and lung cancer, all of the causation that is going on is occurring at the level of particular individuals. Nonetheless, there is clearly a contrast between the *information* conveyed by a claim about the causation of lung cancer in a particular person, Jones and the information conveyed by the claim that smoking causes lung cancer or that smoking caused *n* per cent of the cases of lung cancer in the U.S. population in 2015 or that smoking two packs a day raises the relative risk of lung cancer by such and such. It is this contrast (and nothing more) that I mean to invoke by using the type/token (or “actual cause”) terminology.

Having observed that there is a difference between these two sorts of causal claims, let me note next that the relationship between them is by no means obvious or straightforward – each seems to constrain the other to a much weaker degree than one might naively expect, given that there is a sense in which “underlying” both claims are causal relationships involving the same individuals. First, type-level causal information seems to greatly underdetermine the token or causal attribution relationships holding for particular individuals. If we use lower case letters like *c* and *e* to represent particular events and upper case letters like *C* and *E* to represent types of events of which *c* and *e* are instances, then *C* can be a type cause of *E*, an individual *i* can be subject to events *c* and *e* which are instances of *C* and *E*, and yet it can fail to be the case that *c* causes *e*. For example, even if smoking causes lung cancer, Jones smokes and Jones develops lung cancer, it may not be true that Jones’ smoking caused his lung cancer. Instead his lung cancer may be caused by something else like exposure to asbestos. Similarly, prostate cancer causes (is a cause of) death, but even if Smith has prostate cancer and Smith dies, it does not follow that his prostate cancer caused his death—indeed in older age groups the vast majority of cases of deaths among men with prostate cancer are not caused by prostate cancer. Thus, even if we are given information about type level causal relationships and information about which causes and effects are present in particular individuals, we cannot (in the absence of further assumptions or additional information) infer which are the true causal attribution claims about that individual[[4]](#footnote-4).

When one thinks about it, this should be unsurprising. Establishing that, e.g. a stoke caused Smith’s death requires not just the type level information that strokes cause death and information that Jones had a stroke and died, but additional information as well: For example, assuming (as we are) that Jones’ death is not causally over-determined, it needs to be shown that other type level causes of death either were not present in Smith or, if present, were not “operative” in causing Smith’s death, while the stroke was. This last sort of consideration (which cause was operative) will typically have implications regarding the presence of a process or mechanism by which the cause in question produces death when it does. For example, if Smith’s death is caused by a stroke, this will involve a hemorrhage that impairs the function of brain areas that are vital for life. This causal attribution claim thus has implications for what one might expect to observe in Smith’s brain. Alternatively, if Smith’s death is due to cancer, then tumors will be present that show impairment of the function of vital organs and so on. Often there will be additional signs or symptoms that one of these processes was operative rather than another—e.g. signs of liver failure as a consequence of metastasis of the cancer to the liver. Often information about the time course of the operation of the cause will also be relevant—strokes can produce death very soon after their occurrence; the time course for cancer is typically much slower. Information about type causal relations will not tell us whether or not such particular processes or mechanisms are operative in particular individuals. Indeed many type causal claims (e.g. “smoking causes lung cancer”) don’t provide information about connecting processes or mechanisms at all, even at the type level. To the extent that causal attribution or actual cause information about causes of death claims about processes and mechanisms present in particular individuals and type causal information does not provide such information, it is unsurprising that facts about the latter can underdetermine the former[[5]](#footnote-5).

Moreover under-determination holds in the other direction as well. That is, causal attribution information about the individuals in a population by itself may be insufficient to answer various type level causal questions, particularly about the results of intervention programs, concerning that population, as the following discussion shows.

**3. From Causal Attribution to Predicting the Results of Interventions: The Need for Additional Assumptions**

 To explore this last issue in more detail, let me introduce some background assumptions, some of them obviously unrealistic, which might be used to justify inferences from attribution data to the effects of interventions. Suppose we have a list of all the candidates *C1… Cn* for actual causes of death in a population P and that this list is exhaustive (it contains, in the relevant sense—see below-- all causes of death in P), and that the causes are mutually exclusive so that each death has one and only one cause. Let us call these the *attributable* or *candidate actual causes* of death). Moreover, suppose that we observe, without error, which candidate from this list causes each death in a population[[6]](#footnote-6).

I will make two additional assumptions about this information, both of which seem uncontroversial, in the sense that they reflect generally accepted background empirical facts about the circumstances in which causal attributions of death are made. First, I assume that the “each death has one cause” requirement is understood in such a way that it does *not* rule out each death having in addition more distal causes besides what is cited as “the cause” of death—it is just that these more distal causes do not appear on our list of candidates for actual causes of death. For example, even if an individual death *d* is correctly attributed to a heart attack *h* as its actual cause, I assume this is consistent with *d* having various more distal causes that operate through *h*, so that *h* and hence *d* may have among their causes consumption of a high fat diet and lack of exercise. (Lurking in the background here is the issue of how “proximate” or “distal” the causes that appear on our list of attributable causes of death should be. Why -- or on the basis of what principles -- do we say that heart attack and not the fatty diet or vice-versa should occur on our list of attributable causes of death? This is an important question, but I again propose to put it aside, returning to it in section 7 below.)

 I will also assume although each death *d* is attributed to a single cause *c*, there are usually other factors or conditions besides *c* (call them “background conditions”) that are also causally relevant to *d*— again, although relevant, these don’t appear on the list from which we select “the cause” of death. For example, when a death *d* is attributed to tuberculosis, the state of the subject’s immune system is likely also be causally relevant to *d*, in the sense that some subjects exposed to the tuberculosis bacillus will not develop (or will not die from) TB while other subjects will. However, I take it to be consistent with this that the state of the subject’s immune system is not regarded as among the candidates from which “the actual cause” of death is selected. Both assumptions seem required if we are to interpret the “each death has one cause” requirement in a way that does not make it obviously false.

Given these background assumptions, I now return to the question raised at the outset of this section: Under what conditions does information about the number of deaths attributed to some cause *Ci* in a population allow one to make reliable predictions about what would happen (to death numbers) if one were to intervene to eliminate (or at least reduce the incidence of) causes of this type? For example, suppose we have causal attribution data specifying that some proportion *m* of total deaths within a population are caused by malaria. (That is, *m*= relative frequency of deaths caused by malaria as proportion of total deaths). What if anything can one infer from this information about the reduction in deaths that would occur if one were to successfully intervene to entirely eradicate malaria?

Initially (and naively) it might seem plausible that if the proportion of deaths caused by malaria is *m*, and we eliminate malaria, the proportion or frequency of death in the population (at least in the short term—see below) will decline by *m*. But in fact this will be true only under some very strong additional assumptions.

 To see this let us write:

(3.1) *P(death) = P (death caused by C1) +…+ P (death caused by Cn*[[7]](#footnote-7))

If, as we are assuming, the causes are exhaustive and mutually exclusive, (3.1) is a mathematical identity. In this case we have what Mather et al (2002) call “additive decomposition”, with the different causes of death adding up to 100% of all deaths. Let us also assume that if one successfully eliminates *Ci* , then P(death caused by *Ci* )= 0[[8]](#footnote-8). Finally, and crucially, let us also assume that, in eliminating or removing *Ci*, all of the other probabilities on the right hand side (rhs) of (3.1) remain the same—that is, that the elimination of *Ci* is of such a character that *P(death caused by Cj*) for *i≠ j* is just what it was before the elimination. This is the *independence assumption* mentioned earlier. Then, given this independence assumption, the result of intervening to eliminate *Ci* is just given by dropping the *P (Death caused by Ci) t*erm from the rhs of (3.1) so that the resulting expression is now

 (3.2) *P(death) = P (death caused by C1) +…+ P(death caused by Cj) ..+ P (death caused by Cn*) for all *Cj ≠Ci*

Thus given the independence assumption, the probability of death will be reduced by *P (death caused by Ci)* if we eliminate *Ci* entirely from the population.

Obviously, however, this conclusion depends on whether the independence assumption holds—“depends” in the sense this assumption is sufficient and “almost necessary” [[9]](#footnote-9) for the conclusion. In approaching the issue of the status of this assumption, we should note, as a point of departure, that we need to be more explicit about what is meant by *P(Death*). If *P(Death)* is interpreted as the proportion of individuals in the population who eventually die in the very long run, obviously this is 100 per cent. Under this interpretation, removing one of the causes of death *Ci* must, just as a matter of mathematics, result in an increase in frequency of other causes of death, since these must continue to add up to 100 per cent, thus leading to a violation of the independence assumption. So to give the independence condition even a chance of being true, we must be more restrictive about what is meant by *P(Death).* One natural possibility is to take *P(Death)* to mean death within some specified time interval *Δt* (e.g., within the next 2 years.)[[10]](#footnote-10). Interpreted in this way, it is certainly possible that the independence assumption is approximately true for relatively small values of *Δt* and for some diseases and some age groups in the population For example, it may be that if a young person’s chances of dying from ebola in 2014 were entirely eliminated, that person’s chances of dying from some other disease in 2014-16 would be largely unaffected, at least for many ebola elimination procedures. On the other hand, as we make the time interval *Δt* larger[[11]](#footnote-11) or consider individuals in the population who are older and for this or other reasons more vulnerable to dying from some cause or other (even if they do not die from the particular cause which in fact kills them), it is arguable that the independence assumption becomes less plausible. Moreover, the independence assumption seems more plausible for some causes of death than others. For example, as discussed below, if lung cancer and heart disease are on the list of actual causes of death and many cases of lung cancer are due to smoking, an advance in chemotherapy that eliminates lung cancer may lead to an increase in smoking-induced heart disease as a cause of death. If elderly people suffer from multiple disorders, eliminating one of these may increase the probability that death will be caused by some other disorder.

 To explore in a more systematic way some of the conditions under which the independence condition can fail, let us begin with a non-medical illustration. Suppose there are only two attributable causes of forest fires *F* -- camp fires *C* and lightening strikes *L*, that each *F* is caused either by a *L* or a *C* but not both and that for each *F* that occurs one can establish conclusively whether it was caused by a *C* or an *L* . Subtracting the number of *F*s caused by *C* from the total number of *F*s will not be a good estimate of the number of *F*s that would result if one were to intervene to eliminate or remove all *C*s if (as may well be the case) in the absence of *C*s, the number of *F*s caused by *L* would increase. This might happen because, for example, *C*s burn the forest and eliminate combustible material, decreasing the probability that *L*s will cause fires. Note that this does *not* require that there be a correlation between occurrences of *C*s and occurrences of *L*s - instead what is required is that there be a correlation between whether *C*s *cause* *F*s and whether *L*s *cause* *F*s. In the present scenario this can happen even if *C* and *L* themselves are independent, as long as *C*s eliminate combustible material which is relevant to whether *L*s cause *F*s (and/or vice-versa). Of course it might also be true that *C* and *L* are themselves correlated—for example, because widespread lightening destroys so much forest via fire that people don’t want to go camping or are unable to find enough combustible material to start campfires. In this case too, the correlation between *C* and *L* is likely to lead to a correlation between whether *C*s cause *F*s and whether *L*s cause *F*s.

It seems reasonable to suppose that similar failures of independence might occur for the relationship between different diseases and death. There are a number of different possible scenarios under which this might occur.

Suppose, to enlarge on an example described above, you can reliably observe all the deaths causally attributable to lung cancer within time interval *Δt*. Suppose also that many of the people who died of lung cancer were smokers and that many of these smokers are such that if they had not died of lung cancer within *Δt,* they would have died within *Δt* of heart disease or some other smoking-caused illness. Under such a scenario, to predict the effects of “eliminating deaths from lung cancer” we need to specify more clearly what is meant by the quoted phrase. If eliminating deaths from lung cancer involves new developments in chemotherapy that are effective against lung cancer but not against heart disease, then the incidence of deaths attributable to heart disease may increase if lung cancer is eliminated in the way described, so that the independence assumption is violated. In this scenario, the number of deaths attributed to lung cancer will be an overestimate of the reduction in overall deaths that will result if lung cancer is eliminated[[12]](#footnote-12). On the other hand, if the elimination of lung cancer is the result of an effective campaign that reduces the incidence of smoking, this may also reduce the incidence of other smoking-related illnesses such as heart disease and the deaths they cause, so that independence again is violated. In this case, causal attribution information about the number of deaths due to lung cancer may *underestimate* the effects of lung cancer elimination when this is accomplished by the elimination of smoking removal. This example illustrates a general point to which I will return below: counterfactuals about what would happen if a cause of death were to be removed may be ill-defined or lack clear truth values unless the procedure by which the cause of death is removed is explicitly specified[[13]](#footnote-13). This is particularly likely to be a problem if the cause of death which is envisioned as removed is correlated with or causally related to other causes of death.

In the example just described, the correlation between deaths caused by lung cancer and those caused by heart attack (and the resulting violation of the independence assumption) is due to the presence of a common cause (smoking) of lung cancer and heart attack. However, as noted earlier, violation of the independence assumption can arise even in the absence of anything that we would ordinarily describe as a common cause, just as a consequence of arithmetic and some uncontroversial empirical assumptions. Consider all those people G who die from some type of cause *Ci* in time interval *Δt*. Suppose there are *N0* such deaths. Some number *N1* of those in G would have died from some other cause or causes *Cj  i≠ j* in *Δt* had they had not died of *Ci*. Eliminating *Ci* thus reduces the total number of deaths by *N0* - *N1* which of course is less than *N0* as long as *N1* is greater than zero, which presumably will usually be the case. The shortfall *N1* in death reduction (in comparison with *N0*) will be distributed among each of the other causes of death, presumably (in typical cases) causing many of them to increase. Of course the size of this effect will depend on the details of the case. In some circumstances it may be small enough to be legitimately ignorable. However, the larger *Δ t*, the larger this effect is likely to be. As noted above, the size of the effect will also depend on the age composition and general health of the population. In older populations or populations that are at relatively risk of death from a variety of causes (because of malnutrition, bad sanitation, high levels of violence etc.) the probability that those who die from *Ci* would have died from some other cause in *Δt* may be relatively high.

 Next consider a different example: a disease *D* (e.g., infection with a parasite) that although debilitating is not regarded as a candidate for an attributable cause of death (that is, is not listed as a cause of death on death certificates, is not among the ICD causes of death etc.). Suppose this disease *D* increases susceptibility to other diseases *D\** (within the relevant time interval *Δt*) that do cause death. When a subject with *D* and *D\** dies, *D\** and not *D* will be listed as the cause of death. Going by causal attribution data, since no deaths are attributed to *D*, it will look as though intervening to eliminate *D* will not affect death rates. But it may be that this intervention would have a substantial effect on death rates, in virtue of reducing susceptibility to many death-causing diseases. Alternatively, suppose *D* has a small direct effect on death rates (it occasionally appears on death certificates as a cause of death) but that it also increases the incidence of *D\** which is a major cause of death. Again the impact of *D* judged just from causal attribution data may look small, when in fact removing *D* would have a big impact on death rates. Here again the basic problem is that the relationship between different factors that affect death (*D* and *D\**) is not apparent just from the cause of death attribution data alone.

**4. Causal Modeling: Graphs and Equations**.

The previous section is meant to motivate the more general claim that reliable inferences from (4.1) information about attributions of death to various causes *Ci* to (4.2) conclusions about what would happen to death rates if one were to eliminate those *Ci* requires additional information of various sorts. Either one must have information that an independence assumption of the sort described in Section 3 is satisfied or, if there is reason to think this assumption does not hold, one must model the relationships among the various possible causes of death in the population and the way these will be affected by the contemplated intervention program. Moreover, the causes whose relationships must be so modeled will likely include many causes that do not appear on the list of attribution causes of death one is employing. For example, even if smoking, obesity, and bad sanitation are not on the list of candidates for attribution causes of death (perhaps because they are deemed too distal) relations among them and the correlations they induce among the candidate attribution causes of death will be relevant to predicting the effect of disease intervention programs. More generally, information concerning the causal attribution of deaths of particular individuals to particular causes *Ci* will not by itself tell us what would have happened to those individuals (whether, for example, they would have died from some other cause) if those *Ci* had not caused their deaths[[14]](#footnote-14).

 These observations motivate the following three sections of this paper, which provide an overview of how to think about and model relationships among causes of diseases, cases in which diseases have more than one cause, and how to think about interventions in such situations. I consider two representational devices that can be employed for this purpose: direct graphs and structural equations.

In both the directed graph and structural equation frameworks, causal relata are represented by *variables* where a variable some property or quantity which can take at least two values and perhaps more. Each individual in the population of interest has a value for each variable represented: For example, we might employ a variable *Death* which can take one of two values *= 0* or *1* for each individual, depending on whether that individual dies within some specified time interval and a variable *S* representing how many packs an individual smokes a day (or alternatively whether that individual smokes more or less than one pack per day).

In the directed graph representation, a (direct)[[15]](#footnote-15) causal relationship from variable *X* to variable *Y* (*X* causes *Y*) is represented by an arrow from *X* directed into *Y*:

X Y

We will consider in more detail below how this arrow is to be interpreted but at present let us think of it as encoding (at least) the claim that the value of *Y* is some function of the value of *X*, where the function in question is not a constant function—that is, (some) different values of *X* are associated with different values of *Y*.

 Systems of causal relationships are represented by means of systems or concatenations of arrows:

 X Y Z

*X* causes *Y* which causes *Z*

and

X

 Z *X* and *Y* cause *Z*

Y

One can use directed graphs to represent relationships among causes (which, remember, is what we wanted to be able to do in connection with various causes of death). For example, in this structure:

 X

 Y Z

 *X* causes *Y*, *X* and *Y* cause *Z*. Here *X* might represent the occurrence or not of a debilitating disease which both directly causes death (*Z*) and also causes the occurrence of a second disease *Y* which also causes death.

 An alternative way of representing causal relationships is by means of systems of equations—often called structural equations when interpreted causally. Here each variable representing an effect is written as a function of the variables representing its (direct) causes, following the convention that the former is on the left hand side and the latter on the right hand side of the equation. The equation describes the pattern of dependency of the values of the effect variable on the values of the cause variable. For example, suppose that bites by a rabid dog, if not accompanied by treatment by a vaccine, always cause death, but never cause death with the vaccine. Then if *D= 1* or *0* according to whether subject dies or not, *V= 1* or *0* according to whether subject receives vaccine, and *R= 1* or *0* according to whether the subject is bitten by a rabid dog we have:

*D= 0* if *R= 0*

*D= 0* if *R=1* and *V=1*

*D=1* if *R=1* and *V=0*

This pattern of dependency can be represented more compactly by means of the following equation:

 (4.1) *D= R (R-V)*

Systems of causal relationships can be represented by systems of equations. For example, suppose, as above, that *X* and *Y* (directly) cause *Z* and also that *X* directly causes *Y*. Assume for simplicity the relationships are all linear. Then these causal relationships can be represented as follows:

(4.2) *Z= aX+ bY*

(4.3) *Y=cX*

The first equation (4.1) represents that *X* and *Y* are direct causes of *Z*, and the second (4.2) that *X* directly causes *Y.*

There is a simple relationship between the representation of causal relationships by means of equations and by means of directed graphs. Given an equation

*Z= F(X, Y)*

the associated graph consists of arrows from each of the rhs or cause variables into the lhs or effect variable. Thus the graph associated with

 *Z= F(X, Y)*

 is

X

 Z

Y

The graph associated with

 *Z= aX+ bY*

 *Y=cX*

is

 X

 Z

 Y

Figure 4.1

We should note, though, that there is the following difference between systems of equations and directed graphs: when a system of equations is written down in a way that makes the functional dependencies between the independent and dependent variables explicit, it conveys more information than the associated graph. A graph like Figure 4.1

represents that the value of *Z* depends in some way or other on the values of *X* and *Y*, but does not make it explicit exactly how *Z* depends on *X* and *Y*. An equation does this.

 My discussion so far has said nothing about how graphs and equations that are used to represent causal relationships are supported by evidence or assessed for empirical adequacy. This is a complicated subject about which there is a vast literature. For our purposes, the important points are these: First there are a variety of “causal inference” procedures that infer both directed graphs and systems of equations from the combination of (a) statistical information about the joint probability distribution of the variables in terms of which the graph or equations are characterized *and* (b) additional non-statistical assumptions of various kinds. The latter may be (b1) relatively subject matter specific, reflecting causal background knowledge of various sorts[[16]](#footnote-16) (knowledge that although causal is distinct from the causal conclusions to which one infers). Alternatively (b2) the additional non-statistical assumptions may be relatively domain general, as in the constraint-based approach to causal inference developed by Spirtes, Glymour and Scheines, 2000. Second, and crucially, neither this statistical information (a) nor the non-statistical information (b1 and b2) is information reported in attributions of deaths to individual single causes.

**5. Interventionism**

So far, we’ve been drawing arrows and writing down equations in order to represent causal relationships, but nothing has been said about how the notion of “cause” should be understood. Under what circumstances would it to be correct to draw an arrow from *X* to *Y* (*X🡪Y*) or write down an equation *Y= F(X)* representing that *X* causes *Y*?

One attractive answer to this question draws on the idea that what is distinctive about causal relationships is that they are relationships that are exploitable in principle for manipulation and control. That is: If *X* causes *Y*, then, if you were to manipulate *X* in the right way, *Y* would change and conversely, if for some manipulations of *X*, *Y* changes, then *X* causes *Y*. This can be made slightly more precise as follows:

(**M**) *X* causes *Y* in background circumstances *B* if and only if there is some possible intervention in *B* that changes the value of *X*, such that if that intervention were to occur, the value of *Y* or the probability distribution of *Y* would change– i.e., *X* and *Y* would be correlated under that intervention on *X*.

Before proceeding, some clarificatory remarks may be helpful. First, the causal notion characterized here is what was earlier called a type causal notion. It is used in (**M**) to elucidate claims of the form “*X*s cause *Y*s” but, as we shall see shortly, similar ideas can be used to elucidate the causal claims made by graphs and systems of equations, including quantitative type level causal claims of various sorts such as those employed in epidemiology. Second, the notion of an “intervention” employed in **M** is a technical term that also will receive more attention below. At this point the reader may think of an intervention on *X* with respect to *Y* as idealized unconfounded experimental manipulation of *X* for the purposes of determining whether *X* causes *Y[[17]](#footnote-17)*. That is, we should imagine manipulating *X* in an “ideal experiment” and observing whether *Y* changes under this manipulation. Third, note that (**M)** is a version of what philosophers will recognize as a counterfactual account of causation– or at least it is an account that connects causation and counterfactuals: what matters for whether *X* causes *Y* is what would happen to *Y* if, perhaps contrary to actual fact, an intervention on *X* were to occur. (Thus (**M**) does *not* require that interventions on *X* actually occur for *X* to cause *Y*.) I will call such counterfactuals *interventionist* counterfactuals in what follows and will describe some of the ways such counterfactuals can be evaluated on the basis of non-experimental (that is, “observational” evidence).

(**M**) attempts to capture the idea that whatever else they may involve, causes are (at least in principle) “means” or “handles” for manipulating or changing their effects, in the sense that if it were possible to manipulate the cause this would be a way of manipulating the effect. This is an idea one finds (in various forms) in philosophers such as Collingwood (1937), Gasking (1955), von Wright, and more recently, Price and Menzies (1991) and Woodward (2003). It is also a common idea in the social science literature, in statistics (as represented by the “potential response models” developed by Rubin, 1990 and Holland, 1986), among epidemiologists (Greenland et al., 2008) influenced by Rubin- style ideas and in computer science/machine learning (e.g. Pearl, 2000). Interventionism about causation seems to fit much of causal discourse in epidemiology and public health rather well, since one of the primary reasons these disciplines are interested in the causation of disease, death and disability has to do with getting information that is relevant to manipulation and control of these effects.

As already noted, (**M**) is intended as an account of type-causation and cognate causal notions. It is not intended as (and would clearly be inadequate as) an account of causal attribution or “actual cause” claims. This is so whether or not causal attribution claims should be understood in terms of counterfactuals (an issue that is controversial within philosophy). Even if causal attribution claims can be analyzed or elucidated in terms of counterfactuals, these will be different counterfactuals (they will embody different counterfactual information) than those invoked in (**M**) and used to predict the results of disease interventions. ``

Let us now turn to the issue of how (**M**) might be used to capture what it is for a graph or system of equations to have a causal interpretation or to make true causal claims. Consider a graph consisting just of arrows into a single variable– e.g.

X

 Z

Y

We may think of this diagram as correct (as a representation of causal relationships) as long as there is some intervention on *X* ( for some value of *Y* – in this case, this would be included in the background circumstances *B*) that would change the value of *Z* and similarly for *Y* and *X* interchanged. In the case of a causal structure represented by a single equation– e.g.

(5.1) *Y=aX*,

 the equation is a correct description of the causal relationship between *X* and *Y*, just in case *Y* would change in the way represented in (4.1) under some interventions on *X*. In this case, the equation (4.1) describes the quantitative causal dependence of *Y* on *X*. This general idea needs to be tweaked and extended if it is to be used to capture direct causal relations of the sort represented in directed graphs and equations when these are embedded in larger causal structures but I will put these complications aside here, referring the reader to Woodward, 2003.

 **6. More on Interventions**

 Let me now turn to a discussion that was postponed above. What exactly is meant by talk of an *intervention* on *X* in claims like (**M**)? The notion of an intervention has been extensively discussed in the literature on causal modeling—see, for example, Woodward, 2003, Pearl, 2000, Spirtes, Glymour and Scheines*,* 2000. For present purposes, one may think of an intervention on *X* as an exogenous change in the value of *X* that is appropriately unconfounded from the point of view of inferring whether there is a causal connection from *X* to *Y*. An intervention on *X* (with respect to *Y*) changes the value of *X* in such a way that any change in the value of *Y*, should it occur, occurs only through the change in *X* and not in some other way. Put slightly differently, an intervention on *X* should be such that any change in *Y* can only be due to the change in *X*. Randomized experiments are one paradigmatic example of an intervention but there are many others. A number of techniques for inferring causal relationships from non-experimental data such as instrumental variables and regression discontinuity designs work by attempting to identify intervention-like relationships in such data.

 The effects of interventions have a simple representation in terms of directed graphs and equations. Understanding this will both provide additional motivation for interventionism and will be helpful later when we consider the modeling of disease intervention programs. Consider a so-called common cause structure:

* 1. *E1 🡨 C🡪 E2*

According to this structure, *C* causes *E1* and *E2* but *E1*does not cause *E2* or conversely. *E1* and *E2* are correlated but not causally related. For example, *C* might be smoking, *E1* the presence or absence of yellow fingers, and *E2* the occurrence or not of lung cancer. To intervene on *E1* (in the technical sense of intervention we are trying to elucidate) one must change *E1* in an appropriately exogenous and unconfounded way. This requires, in particular, that *E1* be changed in a way that is independent of *C*, the other cause of *E2*. Examples of such interventions include bleaching the fingers of those subjects who have yellow stains and introducing yellow stains via dye onto the fingers that were not previously yellow, in both cases in a way that is independent of whether such subjects smoke. If (6.1) is correct, under such interventions the correlation between *E1* and *E2* will disappear, reflecting the fact that *E1* does not cause *E2* in accordance with (**M**). In other words, the association between *E1* and *E2* cannot be exploited for purposes of manipulation and control. This shows it to be non-causal.

By contrast, if one intervenes appropriately on *C* (say by reducing the incidence of smoking), the incidence of both yellow fingers and lung cancer will change, showing that *C* does cause *E1* and *E2*. These are relationships that can be exploited for manipulation and control. This illustrates how the interventionist account distinguishes between cause and mere correlation.

 In many cases interventions have a natural graphical representation: one may think of an intervention *I* on a variable *V* as “breaking” all other arrows directed into *V*, putting *V* entirely under the control of the intervention variable, while leaving intact all other arrows in the graph, including all arrows directed out of *V.* For example in the common cause structure,

(6.1) *E1* 🡨 *C*🡪 *E2*.

an intervention *I* on *E1* replaces the structure (6.1) with the following structure

(6.2) *I*🡪 *E1* *C*🡪 *E2*.

(the arrow from *C* into *E1* in (6.1) is “broken” and the intervention *I* is the sole cause of *E1*)[[18]](#footnote-18).

The conception of intervention introduced above is often described as “surgical”– an intervention *I* on *X* with respect to *Y* is supposed to change *X* but not any other variable that is a cause of *Y* and is not on the path or route from *I* to *X* to *Y*. I noted above that some real-life interventions in well-controlled experiments have this feature but many real-life manipulations, including virtually all those aimed at the elimination of diseases in a public health context do not. Instead, these manipulations are almost always “fat-handed” or non-surgical, in the sense that they change other variables besides the variable intervened on and these in turn have independent effects on the dependent variable of interest. An illustration is provided by an example discussed above: in order to reduce the incidence of deaths *D* from lung cancer (*L*), a campaign is introduced that (improbably) reduces the incidence of smoking *S* to zero. The causal structure associated with this manipulation might look something like Figure 6.1 (vastly oversimplified), with *H* representing heart disease and *A* the manipulation associated with the anti-smoking campaign.

 L

A

 S D

 H

Figure 6.1

The manipulation *A* on smoking affects the incidence of lung cancer but it counts as a non-surgical or fat-handed manipulation of lung cancer with respect to death since *A* also affects death via a route that does not go through *L*– namely the route that goes through *H*. Thus the total effect on *D* of reducing *S* will reflect the influence of this reduction both along the route going through *L* *and* the route going through *H*.

 Suppose by contrast we could intervene in a fairly surgical way, just on lung cancer with respect to death– for example, we employ a newly discovered treatment *I* that cures lung cancer without any additional side effects relevant to death. If this treatment is given to everyone with lung cancer, then, by our rules for the representation of interventions this would replace

 L

S D

 H

With

 I L

S D

 H

This amounts to intervening to set the value of lung cancer to zero (that is setting lung cancer = absent) but under this intervention the influence of smoking on heart disease would remain and would still cause death. Thus the overall reduction in deaths would be less than it is in the immediately previous scenario.

 In assuming that the intervention on lung cancer is surgical, we are in effect assuming that the effect of this intervention does not change the probability with which smoking causes heart disease and heart disease causes death. This of course is just a form of the independence assumption used to go from causal attribution data to the results of disease removal that was discussed in Section 3. Put differently, if we assume that elimination of one kind of cause of death via some public health program leaves all other causes of death (and their causal relation to death) unaffected (which is what we assume when we infer directly from data attributing causes of death to predictions of the result of intervention programs on those causes), we are in effect assuming that the elimination program is surgical or intervention-like, an assumption that, as we have noted, will be mistaken in many if not most real life-cases. If the elimination program does not act in a surgical way but instead affects a number of different causes of death via independent routes, we need to carefully model these relationships and their impact on death. Graphs and equations combined with population level information about the incidence and correlations among different causes of death provide one way of doing this.

To further illustrate how this might work, consider again example from Section 3 of the debilitating disease *C* which both directly causes death *D* but also causes death indirectly by making subject more likely to contact another disease *C\** that also causes death. This may be represented as follows:

 C

 C\* D

If *C\** is unlikely to occur if *C* is set equal to zero, then intervening to set *C= 0* is likely to prevent more deaths than intervening on *C\** alone.

 C

 D

 I C\*

Again, note how the diagram can be used to represent this information, while attribution data concerning the attribution of death to *C* and *C\** by itself will not .

**7. Other thoughts/worries about causal attribution data**

 There are other reasons why an exclusive focus on cause of death attribution data in predicting the results of intervention programs can be misleading, particularly when accompanied by the requirement that each death should have just one cause. First, this will often privilege more proximal causes of death over more distal causes. The former are more likely to be listed on death certificates and autopsy reports, presumably in part because their causal role is more likely to be thought to be “observable” at the time of death. (One observes that a heart attack caused the death but not the smoking, fatty diet or the poverty-induced stress that contributed to the heart attack and hence to the death). However, from the point of view of preventing deaths, more distal causes may be the more favorable intervention points and information about the role of distal causes and their interrelations (as well as their relations to more proximal causes) may be crucial in predicting the results of intervention programs.

A second, related point is that there often seems to be an element of convention or arbitrariness in decisions about how “proximal” to death a cause needs to be to count as “the cause” or “the underlying or attributable cause” of death—or at least the methodological rationale for such decisions is often unclear. Welch and Black (2002) describe current death attribution rules according to which a death within a month following surgery for cancer will be attributed to the surgery rather than the cancer, even though the surgery would not have been undertaken if cancer had not been present; they suggest that this rule undercounts (what would naturally be regarded as) deaths from cancer and urge that in such cases the deaths be attributed to cancer. This seems reasonable enough but it also seems implausible that purely medical or biological considerations will single out any single rule as the uniquely appropriate one for dealing with cases of this sort. Similar issues arise when one asks whether diabetes or the heart or kidney failure consequent to the diabetes should be regarded as “the cause” of death.

If one allows considerations having to do with the most effective points of intervention to influence judgments about which are the candidates for attribution causes for death, then it will often be better to focus on more distal candidates. For example, on the assumption that a patient with a severely compromised immune system due to HIV is likely to die soon of something else if she does not die of a particular kind of infection *F,* it would be more informative for HIV to be listed as the cause of death, since reducing the incidence of HIV is likely to be a far more effective way of reducing deaths than trying to act directly to prevent *F*. However, once one decides that considerations related to effectiveness of interventions should influence what is regarded as a cause of death, this may create pressure to locate such causes at even more distal points in the causal chain leading to death--- in the diet and lack of exercise that led to the diabetes, the social policies that made HIV infection more likely and so on, since these may be the most perspicuous intervention points.

In any case, different choices about what to include among the candidates for causes of death, and how proximal or distal to death these are and whether these mainly include conventional diseases and medical conditions or instead include the social, economic or behavioral factors will heavily influence one’s picture of death causation. Moreover, to the extent such information is taken as a basis for prioritizing interventions, it will also influence judgments about the best allocation of medical resources. For example. a recent study lists the following results for what are described as “leading” causes of death in the U.S. in 2000 (in descending order of frequency) : heart disease, malignant neoplasm, cerebrovascular disease, chronic lower respiratory tract disease, unintentional injuries, diabetes mellitus, influenza and pneumonia, Alzheimer disease nephritis, nephrotic syndrome and nephrosis and septicemia In the same paper, another set of causes, described as “actual causes” of death, are listed as follows: tobacco poor diet and physical inactivity, alcohol consumption, microbial agents, toxic agents, motor vehicle crashes, incidents involving firearms, sexual behaviors, and illicit use of drugs Although there is nothing inconsistent about these two lists of causes of death, if causes of death are taken to a basis for prioritizing intervention programs, the two lists may suggest rather different interventions. A causal model of the interrelations among the causes on the two sets of lists would be a useful source of additional information, especially if the goal is to identify the most effective intervention points.

**8. Problems of Variable Selection**

One way of thinking about some of the issues raised in section 7 is that they have to do with *variable selection.* What are (on what basis should one choose) the variables that ought to figure in causal claims? For example, what grounds (if any) are there for focusing on a variable that measures whether a subject has had a heart attack as a candidate for cause of death rather than a variable that instead measures level of obesity? The question of “where do the variables come from?” (or “which are the legitimate or appropriate variables for purposes of causal analysis?”) is an important and under-discussed issue in causal modeling of all kinds, including epidemiology. This is issue is far too complex to fully address here, but here are some brief observations that connect to previous discussions of this problem[[19]](#footnote-19).

Since within an interventionist framework, causal claims are understood as claims about the outcomes of interventions, this naturally suggests two related criteria for variable choice:

(8.1) Variables should be such that they are well-defined targets for interventions

 and

(8.2) Variables should be such that interventions setting them to the same value lead to the same (homogeneous) outcome.

With respect to (8.1), a variable might be such that it is very unclear or underspecified what would be involved in manipulating /intervening on it. Complaints of this sort have been advanced by a number of writers, including some epidemiologists (e.g., Greenland et al., 2008), against the use of certain demographic variables – such as gender or race-- and other variables, such as obesity in causal models. No one doubts of course that such variables can be *correlated* with or predictive of the incidence of diseases and health outcomes[[20]](#footnote-20); the issue instead is whether it is appropriate to think of such variables as playing a *causal* role. If, in order for a variable to figure as a cause, it must be possible in principle to intervene to change it (or at least provide a coherent description of what would be involved in manipulating it) and if, as some claim, it is unclear (perhaps for conceptual reasons) what would be involved in manipulating someone’s race, then race is a defective causal variable.

Alternatively, for some variables, it might be contended that there are a number of distinct possibilities that might be involved in manipulating them and associated with this different possible outcomes under such manipulations (non-homogeneity of outcome). This contention has been advanced recently by Hernan and Taubman (2008) regarding obesity (as measured by body mass index) as a cause of shortened life. The authors argue that claims about the effects of obesity on life expectancy are defective because (i) they do not explicitly specify the interventions on body mass index (BMI) that are being considered and (ii) different methods for modifying BMI may lead to different mortality outcomes, even if each of these involves setting BMI to the same value in a given person. For example, a reduction in BMI as a result of increased exercise may have a rather different impact on life expectancy than a reduction in BMI due to reduction of caloric intake and a reduction of BMI due to bariatric surgery is likely to have a still different impact. It would be better to replace obesity or BMI as a candidate causal variable with more fine-grained variables that specify more exactly what is involved in manipulating BMI.

Greenland (2002), in the course of criticizing Mathers et al. (2002), appeals to related considerations in arguing that causal effects should be ascribed only to *actions*– e.g. to an anti-smoking campaign (presumably implemented in a specific way)– rather than to *results* of actions (e.g., a diminishment in the level of smoking by a certain amount). Greenland’s motivation, as I understand it, is that when one specifies an action (rather than just a result) this ensures that the intervention or manipulation associated with the causal claim is made explicit. Although (as is obvious from my discussion above) I agree with Greenland that some variables are unsuitable candidates for causes (because it is unclear or ambiguous what is involved in manipulating them), I think that Greenland’s proposal is too restrictive. On my view, there is nothing wrong with thinking of variables that do not correspond to actions as causes (science is full of causal claims of this sort—e.g., claims involving variables like “mass” and “charge”. What is problematic are variables that are not clear targets for manipulation or that, depending on how their values are brought about or realized, appear to lead to different effects (or involve inconsistent counterfactual claims about what would happen if those variables were to take those values.) Variables like “gender” and “obesity” arguably have this feature, but it is possible to define non-action variables whose effects on some target do not depend on how they are realized.

 **9. Counterfactuals**

As should be obvious from the preceding discussion, my view is that the use of counterfactuals and counterfactual reasoning is essential to reasoning about causation and the effects of intervention programs. It is simply not possible to avoid appealing to counterfactuals (or something equivalent) in thinking about these issues. However, the use of counterfactuals that are unclear or ambiguous can create lots of problems. A theme running through this paper (and portions of the literature in epidemiology and elsewhere) is that in using counterfactuals it is very important to make their antecedents as explicit as possible. A common default, in using a counterfactual to convey causal information (as in, “If lung cancer were eradicated, the short term death rate would decline by such and such”) is to interpret the counterfactual as though its antecedent is realized by a surgical intervention, in the sense of intervention characterized above. (This is the interpretation that is in effect adopted in by those who conceptualize causation in terms of “potential response”, both in statistics and in econometrics and epidemiology[[21]](#footnote-21).) As should be obvious from my discussion above, I believe that this interpretation has much to recommend it, if we are interested in connecting causal claims to counterfactuals, but it is also important to recognize that such surgical interventions may not correspond to any actual manipulation of disease incidence that we might perform. Instead, virtually all real- life programs aimed at affecting causes of disease and death will be non-surgical and fat-handed, as in the case of an anti-smoking program aimed at reducing the incidence of lung cancer, which also reduces heart disease as well. Thus when we make a counterfactual claim about what would happen to death rates if we were to successfully reduce the incidence of lung cancer, we need to make it explicit just what possibility we have in mind in connection with the antecedent of this counterfactual (do we mean a therapy that affects only lung cancer and meeting the technical requirements for an intervention, or do we mean a reduction that occurs via an anti-smoking campaign?).

 Problems raised by the use of counterfactuals with unclear antecedents are particularly pressing in connection with counterfactuals whose antecedents involve claims about the *absence* or *non-occurrence* of some factor— and of course such counterfactuals are used very frequently in the explication of causal claims. The basic difficulty is that there are many different ways in which some factor *C* may fail to occur and an antecedent like “If *C* had not occurred…” fails to explicitly specify which of these is intended. Sometimes the context will resolve any ambiguity but when this is not the case, the solution again is to make the intended antecedent explicit. For example, one might replace. “if Jones had not smoked three packs a day…” (which leaves it open whether this antecedent is realized by Jones’ smoking 2.5 packs or none at all) with a determinate antecedent specifying Jones level of smoking in the counterfactual scenario being envisioned. Similarly, in a public health or epidemiological context, rather than employing counterfactuals like “if malaria did not occur, then…” it would be much better to employ counterfactuals that make it explicit just what particular possibility is being envisioned under this antecedent. This would include, among other considerations, a specification of just how the non-occurrence of malaria is envisioned as being brought about. Alternatively, if bringing about the complete non-occurrence of malaria is not regarded as feasible, one can consider various counterfactual scenarios in which the incidence of malaria is reduced by specified amounts.

 **10. Attributable (excess) fractions and etiological fractions**

I conclude by connecting the material in this paper to the contrast between “attributable” (or “excess”) and “etiological” fractions developed in a series of papers by Greenland and Robins (e.g., 1988, among others). As we shall see, this contrast echoes features of the distinction between population level type causal claims and attributions of causes to particular outcomes and the difficulties of inferring from one of these to the other emphasized above.

Following Greenland and Robins, suppose that we identify a group G that is exposed *E* to some cause of disease *D* and also a perfect non-exposed control group that is exactly like G with respect to all other factors besides *E* that are relevant to *D*. Thus we assume that standard worries about confounding are eliminated. Let us define the *attributable* or *excess* fraction as the fraction of cases of disease in G that would not have occurred by time *t* had exposure not occurred (these are the “excess” cases). Note that we need to make reference to a time *t* in order to define the excess fraction, since this will vary depending on *t*. Assuming no confounding, the excess fraction can be found by comparing the number of cases in the exposed group with the number of cases in the non-exposed group. Define the *etiological* fraction as the fraction of cases in G that are *caused* by the exposure *E*. Greenland and Robins say that the etiological fraction can be defined without reference to time, but I think a better way of putting matters is that, for many diseases there is, as an empirical matter, some specifiable time by which exposure will cause the disease if it does at all—a time that thus can be associated with the etiological fraction in a non-arbitrary way. For example, an exposure to rabies will kill you by a certain time interval after exposure if it kills you at all. By contrast, the investigator has a choice of the time interval associated with the excess fraction, which is not dictated by the nature of the disease, so that in this case the choice is somewhat arbitrary.

As Greenland and Robins show, the excess and etiological fraction are not just conceptually different but can be arbitrarily far apart. To take one of their examples, consider exposure to rabies in the mid-nineteenth century when it was always fatal—here “exposure” is exposure to rabies and the “disease” is death. Among those who are exposed to rabies and die, the etiological fraction is presumably very close to 1—the great majority of the deaths in the exposed group (within the interval in which rabies can cause death) are caused by rabies. On the other hand, if *t* is taken to 120 years, the excess fraction is presumably zero since all of those in G will be dead after 120 years even in the absence of exposure.

As another example, consider exposure to ionizing radiation and the disease of leukemia. As before, let the time for determination of the excess fraction be *t*. Consider the following three types among those who have been exposed and develop the disease (“cases”). Here I quote from Greenland and Robins:

Type 0: The exposure had no impact whatsoever on the case's incidence time. Type 1: The exposure made the case's incidence time earlier than it would have been in the absence of exposure (so exposure played a role in the etiology of this case), but had exposure never occurred (or had its effect been blocked), this subject would still have become a case by t, although later in the interval.

 Type 2: Had exposure never occurred, the subject would not have become a case by t because, in the absence of exposure, disease would have occurred after t, or not at all. Let the number or set of each of these types be denoted *Ao*, *A1*, and *A2,* respectively, with A+ = *Ao*, + *A1* + *A2* and let *M* equal the total number of cases. (1988, p. 1186)

Then the excess fraction is *A2*/*M* and the etiological fraction is (*A1* + *A2*)/*M*. Thus the etiological fraction is always larger than or equal to the excess fraction.

 To connect this distinction with the material in previous sections, suppose we could somehow identify all of the cases of leukemia that were caused by exposure. (This would be the analog of identifying each cause of death via causal attribution procedures. In the case under discussion this would mean that one could somehow establish, perhaps through pathological or forensic investigation, for each case of leukemia whether it was caused by the exposure of interest). This number would reflect the etiological fraction. If we took this to also be equal to the excess fraction we would *over-estimate* the latter (recall that the excess fraction underestimates the etiological fraction). This would be a version of the mistake discussed in Section 3, when one takes the number of deaths caused by a disease, as identified in attribution data, as an estimate of the number of deaths that would not occur if the disease were to be eliminated. In the leukemia example as described this use of attribution data would overestimate the number of cases that would be eliminated. If one is interested in the number of deaths that would be eliminated by time *t* by (surgically) eliminating a disease, the excess fraction (or something like it), interpreted causally as it is above, with any possibility of confounding being ruled out, seems to be the appropriate statistic to use and not the etiological fraction, which, as I have said, would reflect the results of causal attribution data.

Greenland and Robins emphasize the impossibility of identifying the etiological fraction from the excess fraction (or from population level statistics more generally) without strong additional biological assumptions about the details of the disease mechanism in the individuals affected, information that is not reflected in the population-level data to which epidemiologists have access. A similar point seems to hold for the problem of going from information about the etiological fraction to information about the excess fraction, which is (I am suggesting) roughly what the problem of going from causal attribution data to claims about the results of disease elimination programs amounts to: it seems equally unsolvable without strong additional information, either about independence or disease mechanisms.

**References**

Collingwood, R.G. “On the So-Called Idea of Causation”. *Proceedings of the Aristotelian Society* 1937; **38**:85-112.

Gasking, D. “Causation and Recipes”. *Mind* 1955; **64**: 479-87.

Greenland, S. (2002) “Causality Theory for Policy Uses of Epidemiological Measures”, in Murray, C., Salomon, J. Mathers, C. and Lopez, A. (eds) *Summary Measures of Population Health: Concepts, Ethics, Measurement and Applications*. Geneva: World Health Organization, 291-302.

Greenland, S. and Robins, J. (1998) “Conceptual Problems in the Definition and Interpretation of Attributable Fractions” *American Journal of Epidemiology* 128: 1185-1197.

Greenland, S., Rothman, K. and Lash, T. (2008) “Measures of Effect and Measures of Association” in Rothman, K., Greenland, S. and Lash, T. (eds.) (2008), 51-70.

Halpern and Pearl (2005) “Causes and Explanations: A Structural Model Approach. Part 1: Causes” *British Journal for the Philosophy of Science* 56: 843-57.

Hernan, M. and Taubman, S. (2008) “Does obesity shorten life? The importance of well-defined interventions to answer causal questions” *International Journal of Obesity*  **32**, S8–S14.

Hitchcock, C. (2001) “The intransitivity of Causation revealed in Equations and Graphs” *Journal of Philosophy* 98: 273-99.

Holland, P. (1986) “Statistics and Causal Inference*” Journal of the American Statistical Association* 81: 945-60.

Lewis, D. (1973) “Causation” *The Journal of Philosophy* 70: 556-567.

Lewis, D. (1973) *Counterfactuals.* Cambridge, MA: Harvard University Press.

Mathers, C., Ezzati, M., Lopez, A. Murray, C. and Rodgers, A. (2002) “Causal Decomposition of Summary Measures of Population Health” in Murray, C., Salomon, J. Mathers, C. and Lopez, A. (eds) *Summary Measures of Population Health: Concepts, Ethics, Measurement and Applications*. Geneva: World Health Organization 2002, 273-290.

Pearl, J. (2000) *Causality: Models, Reasoning, and Inference*. Cambridge: Cambridge University Press.

 Price, H. and Menzies, P. “Agency and Probabilistic Causality”. *British Journal for the Philosophy of Science*.1991; **42**:157-76.

Rothman, K., Greenland, S. and Lash, T. (2008) *Modern Epidemiology*. Philadelphia: Lippincott, Williams and Wilkins.

Rubin, D. “Estimating Causal Effects of Treatments in Randomized and Nonrandomized Studies”, [*Journal of Educational Psychology*](http://en.wikipedia.org/wiki/Journal_of_Educational_Psychology)*.* 1974; **66**: 688-701.

Rubin, D. (1990) “Comment: Neyman (1923) and Causal Inference in Experiments and Observational Studies” *Statistical Science* 5: 472-480.

Spirtes, P. Glymour, C. and Scheines, R. (2000) . *Causation, Prediction, and Search*. Cambridge, MA: MIT Press.

Spirtes, P. and Scheines, R. (2004) “Causal Inference of Ambiguous Manipulations” *Philosophy of Science* 71: 833-845.

Welch, G. and Black, W. (2002) “Are Deaths Within 1 Month of Cancer-Directed Surgery Attributed to Cancer?” *Journal of the National Cancer Institute*, 94, 1066-1070.

Woodward, J.(2003) *Making Things Happen: A Theory of Causal Explanation.* New York: Oxford University Press.

Woodward, J. (2006) “Causal Models in the Social Sciences” in Turner and Risjord, eds. *Handbook of the Philosophy of Science, Volume 8 (Philosophy of Anthropology and Sociology)* Elsevier, 157-210.

Woodward, J. (2016) “The Problem of Variable Choice” 193: 1047-1072.

Woodward, J. 2018) “Causal Cognition: Physical Connections, Proportionality, and the Role of Normative Theory” in W. Gonzalez (ed.) *Philosophy of Psychology: Causality and Psychological Subject* Berlin: de Gruyter.

1. The contrast between causal attribution and counterfactual analysis is potentially misleading in another respect as well. A robust tradition in philosophy and elsewhere (e.g., Lewis, 1973, Halpern and Pearl, 2005) claims that causal attribution claims (or as they are also often called, “actual cause” claims) can be analyzed or elucidated in terms of counterfactuals. However, as discussed below, the counterfactuals employed for this purpose are different from the counterfactuals associated with the population-level causal claims that figure in what Mather et al. call “causal analysis”. From this perspective, the relevant contrast is between two different kinds of counterfactual claims, rather than between claims involving counterfactual analysis and causal attribution claims that are not associated with counterfactuals. [↑](#footnote-ref-1)
2. For example, one way of defining the notion of excess fraction is: *EF= RE- RU/RE.* If *RE* is just taken to be the risk (or proportion of cases) among those exposed and *RU* the risk among those not exposed (with no implication that *RE* measures the risk *caused* by exposure or what would happen if the exposure were surgically removed), then *EF* is merely a measure of association. On this understanding, a high value for *EF* is consistent with the exposure playing no role at all in causing the cases (because, e.g., the exposure is merely correlated with some factor that does cause the cases—that is, confounding is present.) On a causal interpretation of *EF*, *RE* is instead as understood as the risk caused by or due to or (causally) attributable to the exposure or perhaps what the risk would be if the exposure were surgically removed. Here the assumption is that confounding has been adequately controlled for, via modeling of confounders or in some other way—e.g. the use of a randomized experiment. Similar remarks apply to the notion of relative risk. When interpreted causally, excess fraction, relative risk and so on are legitimate (type) causal notions—my point is that they do not provide information about actual causes of particular deaths or disease occurrences. [↑](#footnote-ref-2)
3. When I say that type level causal claims in epidemiology are *supported* by statistical evidence concerning correlations I mean just that. On my view, causal claims, whether type or token, are not reducible to or fully analyzable in terms of claims about correlations or associations—see Section 5. Inference to causal conclusions involves combining correlational information with other sorts of assumptions, rather than inference from correlational information alone. [↑](#footnote-ref-3)
4. Some philosophers and others hold that token causes sometimes operate probabilistically. To the extent, this is so, the problem of inferring from type to token causal conclusions becomes even more difficult that it would otherwise be. I don’t think that anything that follows in my discussion depends on whether token causes are assumed to operate deterministically or not. [↑](#footnote-ref-4)
5. As I discuss in more detail below, this under-determination is related to the distinction made in the epidemiological literature between etiological and excess fractions and the under-determination of the former by the latter. In general, determining whether some exposure was an actual cause of a particular death (or how many deaths are causally attributable to some particular cause of death) requires rather specific “biologic” information that goes beyond what is ordinarily available at the type level. [↑](#footnote-ref-5)
6. Obviously this is a highly unrealistic assumption since there is a great deal of evidence that recorded cause of death information is unreliable to varying degrees, in some populations, causes for many deaths are not reported, imputed causes of death based on “verbal autopsies” or other sorts of information can be unreliable etc. [↑](#footnote-ref-6)
7. Note that *P(Death caused by Ci) ≠ P(Death. Ci ).* *P(Death. Ci )* is the probability that *Ci* is present in anindividual in the population and that this individual dies. As observed above, it is entirely possible for *Ci* to be present in an individual or a group of individuals, for those individuals to die and yet for the cause of death in those individuals to be something other than *C*i. Quantities like *P (Death.Ci )* and *P (Death/Ci)* can be estimated from population level statistics (the joint distribution) involving these variables. By contrast, as argued below and as observed by many epidemiologists (e.g. , Greenland and Robins, 1988) quantities like *P (Death caused by Ci)*  cannot be identified from population level statistics or the joint population distribution without very strong additional assumptions of a “biologic” nature about the details of the mechanisms by which the cause in question produces death. [↑](#footnote-ref-7)
8. This assumes away complications having to do with induction periods for different diseases. Even if we eliminate *Ci* at time *t0*, exposure to *Ci* at earlier times may still cause death after *t0* but within whatever interval *Δt* we using to measure deaths due to *Ci* (see below). In such cases, elimination of *Ci* at *to* will not reduce deaths due to *Ci* that occurafter *to* to zero, contrary to what we have assumed. We might try to avoid this problem by requiring that the time at which we evaluate the reduction in *Pr (Death)* due to removal of *Ci* at *to* is sufficiently long after *to* that all deaths due to exposure to *Ci* before *to* have occurred. However, this creates new problems since, as observed below, the longer the interval *Δt* between *to* and the time at which death reduction is evaluated, the larger departures from the independence assumption are likely to be. [↑](#footnote-ref-8)
9. “Almost necessary” in the following sense: It is possible in principle that elimination of some *Ci*might increase the number of deaths caused by some second cause *Cj* and decrease the number of deaths caused death by some third cause *Ck* in such a way that the changes due to *Cj* and *Ck* exactly balance, so that that the total number of deaths eliminated by removing *Ci* is just the deaths attributed to *Ci* even though independence is violated. I assume, however, that such exact balancing scenarios are rare, so that usually satisfaction of the independence condition will be necessary for (3.2) to hold. [↑](#footnote-ref-9)
10. [↑](#footnote-ref-10)
11. The relationship between *Δt* and satisfaction of the independence condition will also depend upon considerations having to do with induction period for various death causing diseases in the population. I will return to this consideration below. [↑](#footnote-ref-11)
12. One way of thinking about this example is as follows: the occurrence of a disease or of a death caused by the disease can furnish information about an underlying condition of the person subject to the disease. This underlying condition may be such that it increases the probability (in comparison with those who lack the condition) that some other cause will produce death. In such cases independence may be violated: the death of Smith from some infectious disease *D* may be evidence that his immune system is weak, in which case if *D* is eliminated from the population Smith may be more likely to die from some other infectious disease. [↑](#footnote-ref-12)
13. This point is emphasized by a number of epidemiologists—see, e.g., Greenland (2002). [↑](#footnote-ref-13)
14. There is a tendency among some metaphysically oriented philosophers to assume that actual (or token or singular) causal claims are in some way fundamental or primary, with type level claims being in some way derivative from or parasitic on these, perhaps because they are generalizations over token level claims. (Thus it might be claimed that the true token level claims, taken in their entirety, “ground” the type level claims or “fix” their truth values or that the latter supervene on the former.) I take no stand in this paper on this “in principle” metaphysical claim. My concern is instead with what follows from the known or practically knowable causal attribution information—that is, with the claims about actual causes of death that researchers are able to establish on the basis of the evidence available to them such as death certificates and autopsy reports. My contention is that *this information* underdetermines the truth value of various type level causal claims about what would happen under intervention programs. [↑](#footnote-ref-14)
15. See Woodward, 2003 for discussion of the notion of a direct cause. [↑](#footnote-ref-15)
16. For example, type –causal background knowledge that tells us that a certain factor *X* is just not capable of causing some effect *Y* may, when combined with information about a correlation between a third factor *X* and *Y* lead via an eliminative argument to the conclusion that *Z* causes *Y* (other factors like X being eliminated from this role.) For additional discussion, see Woodward, 2006. [↑](#footnote-ref-16)
17. “Intervention” is thus a causal notion which in turn is used to characterize what it is for *X* to cause *Y*. For arguments that the resulting characterization is nonetheless illuminating, see Woodward, 2003. [↑](#footnote-ref-17)
18. Interventions also have a natural representation in systems of equations—see Woodward, 2003. [↑](#footnote-ref-18)
19. For some additional discussion, see Woodward, 2016. [↑](#footnote-ref-19)
20. For example, certain diseases may cause death more frequently among members of one gender or certain racial groups. [↑](#footnote-ref-20)
21. With the added feature that the potential response framework appeals to two different counterfactuals, corresponding to two different levels of the cause or exposure variable, but each given a basically interventionist interpretation. [↑](#footnote-ref-21)