2

Frameworks for Causal Inference in Epidemiology

Raquel Lucas
Department of Clinical Epidemiology, Predictive Medicine and Public Health, University of Porto Medical School and Institute of Public Heath of the University of Porto, Portugal

1. Introduction

In 1884, Robert Luedeking, Professor at the St. Louis Medical College and member of the St. Louis Board of Health, published a paper entitled *The chief local factors in the causation of disease and death* (Luedeking 1884), in which he wrote the following:

[In St. Louis, in 1883, the population density of 9.8 persons to the acre] is indeed a low density compared with that of most metropolitan cities: that of London, for instance, is given at 52.5 to the acre in 1883. And yet we find the annual rate of mortality per thousand in London in 1883 to have been but 20.4, while that of St. Louis was 21.35. With such a variance existing in the relative densities, it must needs force itself upon our conviction that inherent faults in our sanitation must be the cause.

In this paper Luedeking compares crude death rates between cities and finds that mortality in St. Louis is slightly higher than in London, even though population density is substantially lower in the former city. He then implicitly uses previous knowledge to attribute the unexpected similarity in mortality, given very different population densities, to deficient sanitation in St. Louis. This paragraph is illustrative of a process in which the application of causal inference to the improvement of population health is attempted: observing an unexpected difference (or a surprising similarity), identifying a cause based on observed data and expert knowledge, and recommending a public health action.

This provides an interesting example of a pragmatic concept of cause in epidemiology. Today, health authorities would probably avoid such strong causal statements. However, it seems unfair to neglect that improving sanitation in St. Louis would very likely decrease mortality substantially at the time.

A cause can be defined as a person or thing that acts, happens, or exists in such a way that some specific thing happens as a result; the producer of an effect (Dictionary.com 2011). On the one hand, this definition reflects the notion that causation is an essential component of the human understanding and interaction with the world. On the other hand, although this seems like a straightforward definition, which is probably in agreement with many if not most individuals' concept of cause, it raises a number of questions: Does the cause always produce the effect? Are other causes involved in producing the effect? If the cause was

removed would the effect be produced? Formal discussions of these and other questions have been the focus of philosophical and scientific approaches to causation and the appropriate notion of cause for a certain discipline is influenced by the kind of causal knowledge that the discipline aims at producing.

While theories on causation and causal inference have been abundantly discussed in the philosophical literature, an operational concept of causation is essential to conduct scientific research. In fact, science has developed experimentation as a useful approach to dealing with the complexity of causal inference. In etiologic epidemiology we are interested in understanding the mechanisms of disease causation and we aim at identifying targets for intervention in order to be ultimately able to reduce the burden and consequences of disease in the population.

In epidemiologic research, however, the multifactorial etiology of human disease, the non-modifiable nature of a number of health-related factors, and mainly the vastly unethical nature of experimentation in human subjects brought about the need to design and conduct studies that are knowingly imperfect approximations to the experimental ideal. In fact, in etiologic studies of human disease we are faced with a number of problems that threaten causal inference and whose avoidance and discussion are at the core of epidemiologic research.

In practical terms, stating that causality is the best explanation for an observed association is equivalent to ruling out, with reasonable confidence, alternative explanations such as reverse causation, selection bias, information bias, confounding and chance. The formalization and discussion of these alternative explanations has become in fact so important in epidemiologic research that it was pointed out that these methodologic issues became the main focus of epidemiology textbooks, at the expense of little attention devoted to the discussion of such fundamental issues as theories of causation or hypothesis formulation (Krieger 1994). Indeed, there has been a shift in recent years towards framing the thinking and teaching of epidemiologic methods into a more solid theoretical basis for causal inference (Rothman et al 2008).

Even though causal inference is such a central issue in epidemiology, and perhaps because of that, different views on causation have proliferated in the epidemiologic literature. A systematic review of scientific publications (Parascandola & Weed 2001) has identified several different explicit or implicit definitions of cause within the epidemiological literature, which the authors classified in the following categories:

- production: causes are seen as part of the production of disease;
- necessary causes: causes are conditions without which the effect cannot occur;
- sufficient/component causes: one sufficient cause guarantees that the effect will occur, and each sufficient cause is made up of component causes, none of which is enough to produce the effect;
- probabilistic causes: causes are seen as conditions which increase the probability of an effect, regardless of whether or not they are necessary or sufficient;
- counterfactual causes: the presence of a cause, compared with its absence, makes a difference in the occurrence of the outcome, while all else is held constant.

While no single model can aspire to provide the answer to causal questions in epidemiology, inferring causation from observed data in human populations is a complex

task which extends beyond the discussion of systematic or random errors, some of which may be dealt with through statistical methods.

The aim of this chapter is to provide a brief overview of selected frameworks frequently used to assist causal inference in epidemiology. Although there are many interesting approaches to causation in the epidemiologic literature, the ones referred below were chosen for historical significance or because of their increasing relevance in epidemiologic research. An additional interesting aspect of the approaches chosen is that they originate from different areas of knowledge, from philosophy (the sufficient and component causes model and the counterfactual model) to medicine/biology (Hill's considerations) and computer science (causal diagrams).

2. Frameworks for causal inference

2.1 Bradford Hill's considerations regarding causation

During the first half of the 20th century it became increasingly clear that monocausal theories of human disease were virtually useless to explain chronic conditions (Broadbent 2009). Models that were previously applied to explain the distribution of disease proved useful, to a large extent, to prevent or treat communicable conditions. However, they were clearly insufficient to uncover the multifactorial etiology of complex chronic diseases and therefore inadequate to identify targets for chronic disease prevention. The increasing interest in understanding the etiology of non-communicable diseases brought about the challenge of re-thinking the process of causal inference from observed data.

In 1965, in his famous paper entitled *The environment and disease: association or causation?*, Sir Austin Bradford Hill recognized the fundamental problem of deriving a causal interpretation from observed associations between exposure and disease (Hill 1965). In this widely known paper, his aim was to provide guidance regarding the process that goes from finding an association between an exposure and a disease to deciding that causation is the most likely explanation for that association. His approach clearly intends a detachment from the philosophical discussion of causation. Rather than proposing a theoretical model for causal inference, he puts forward a set of empirical aspects of associations that should be examined to guide the judgment of causality, namely:

a. **Strength**: this consideration is based on the premise that the stronger an association is the less likely it is that there is some alternative unknown explanation rather than causation;
b. **Consistency**: this refers to the replication of the findings in different methodological, geographical and time settings;
c. **Specificity**: this refers to the high probability that an exposure is causally linked to some outcomes more than to others;
d. **Temporality**: this means that the putative cause must precede the effect and is probably the only indisputable criterion for causality;
e. **Biologic gradient**: this refers to the existence of a dose-response relation by which increased dose of exposure is related with increased expression of the outcome;
f. **Plausibility**: this refers to the agreement of the examined association with existing biological knowledge;
g. **Coherence**: this argues for the importance that the association is not conflicting with scientific knowledge on the disease;

h. **Experiment**: this refers to the possibility of eliciting the outcome by experimentally introducing exposure or, in human populations, an alternative such as the possibility of preventing the outcome by removing the exposure;
i. **Analogy**: this refers to the existence of similar outcomes with exposures of the same kind.

It is noteworthy that the ranking of these considerations was intentional. As acknowledged by Hill, none of these items is necessary for deciding that an association translates a causal relation (except for temporality), and the whole set is not sufficient to prove causation. Therefore, the author himself clearly states that this should not be used as a checklist for causal inference, though it has thereafter frequently served that purpose. Such widespread misuse has probably exposed the limitations of this set of considerations and they have often been dismissed as having little utility in causality assessment. However, it should be noted that one of the major contributions of Hill's paper is probably not only the list of viewpoints but also his reflection that, in the presence of an association, the fundamental question that the practice of epidemiology attempts at answering is: "is there any other of explaining the set of facts before us, is there any other answer equally, or more, likely than cause and effect?". In fact, the key notion is that, if a more likely explanation for the observed association exists, it will probably emerge from the analysis of one or more of Hill's considerations. Another important contribution of this paper is the clear statement that significance testing is useful to quantify the magnitude of the role of chance but adds nothing to the purpose of deciding whether or not an association is causal.

Aiming at analyzing the role of the infection with the human papilloma virus as a potential necessary cause for cervical cancer, Bosch et al underwent the massive process of reviewing available evidence and formally assessing the concordance of observed data with the considerations published by Hill, as well as those subsequently adopted by the International Agency for Research on Cancer (Bosch et al 2002). In this extensive review, by combining those considerations with the sufficient and component causes model, the authors conclude that the role of the infection is consistent with that of a necessary cause for cancer. Although compelling evidence of a causal relation had been available for a long time, their work takes a formal approach to ruling out alternative explanations and is additionally relevant in showing the role of Hill's important considerations in current causal inference.

2.2 Sufficient and component causes model

Aiming at bringing together a philosophical view of cause and the practice of epidemiology, Kenneth Rothman proposed an application of a sufficient and component causes model to epidemiology (Rothman 1976). Central to this model are the notions that each person is susceptible to multiple diseases and that each disease is a multifactorial outcome that results from the co-occurrence of several factors. The minimum set of causes enough to elicit disease is called a **sufficient cause**. Since it is reasonable to assume that not all individuals develop the disease through the same causal process, different sufficient causes are possible for the same disease, each with a different combination of factors (Figure 1, sufficient causes I to IV). Each of the factors that build up one or more of the sufficient causes is called a **component cause** (Fig. 1, CC_1 to CC_7). One of the interesting features of this model is that it enables the representation of sets of component causes that are assumed to intervene in disease causation but have not yet been identified (component causes U in Figure 1).

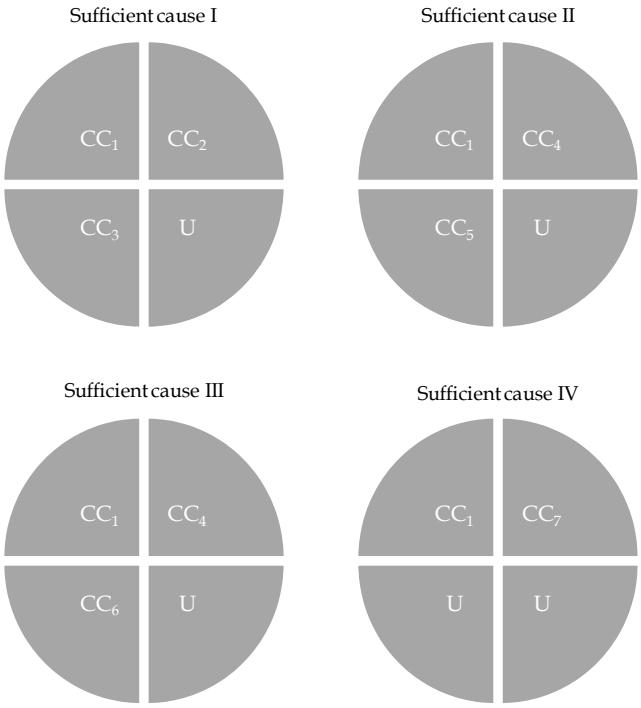


Fig. 1. Hypothetical examples of four sufficient causes for a disease. CC_1 to CC_7 – known component causes; U – unknown component cause(s)

In order to have a causal interpretation, component causes must be defined relative to a clearly defined contrast, i.e. a reference state with which the index state is compared. Since the underlying reality from which causes emerge may have various patent or latent dimensions, each of which may be seen as a continuum, component causes should be defined in relation to a number of attributes. These attributes are essential to conceptualize the presence of each component cause relative to the above-mentioned reference state. To clarify this, we may consider an example where we are studying the occurrence of an event such as venous thromboembolism in women as a possible adverse reaction to the use of a particular, clearly identified hormonal preparation. To define the use of that specific medication as a component cause, we should specify attributes such as **dose** (What mass of drug per day does the component cause refer to? Is the dose defined in relation to body size? Is the cumulative amount of life exposure a relevant attribute?), **duration** (How much exposure time should be considered in the index state? How is intermittent use of the drug relevant to the cause? Is there a particularly decisive biological timing of exposure, such as a specified gynecological age?), **induction period** (Which is the relevant time interval between the occurrence of the first component cause in the specific sufficient cause(s) of thromboembolism in which the hormone preparation is included and the completion of that sufficient cause?), and **reversibility** (Is it reasonable to assume a transient effect of hormone preparations on thromboembolism? Can the elimination of that exposure be considered complete, i.e. is a woman who never used the preparation in the same circumstances relative

to the presence of the component cause as a woman who has never used it?). Although these examples of attributes are by no means exhaustive, they illustrate a part of the complexity in the decision of what constitutes a component cause. At the same time they elucidate on the need for its definition as clearly as possible when discussing causation.

Additionally, component causes may be **necessary** for the occurrence of disease if they are present in every possible sufficient cause (such as CC_1). Disease occurs in an individual when one of the sufficient causes is completed, i.e. all its component causes are present. The period throughout which component causes accumulate to finally produce clinical disease is called the **latent period**.

It should be noted that different component and therefore sufficient causes have different frequencies in the population. The notion of risk at the individual level can therefore be translated into whether or not a sufficient cause is completed. At the group level risk becomes the proportion of people in whom a sufficient cause is completed. An implication of this reasoning to observed data is that the strength of association of a component cause with the occurrence of disease is directly dependent on the frequency of the other causes with which it shares a sufficient cause.

It also results from this model that if one component cause may be identified and prevented, all cases of disease that result from the sufficient cause or causes in which that component cause is present will be avoided. This is in accordance with the intuitive notion that identifying and subsequently eliminating a necessary cause will eradicate disease since it will impair the completion of all sufficient causes. In terms of measures of impact, this model implies that the etiologic fraction of a component cause is the proportion of disease that is attributable to the sufficient causes that contain that component.

This model provides a clear conceptual meaning for effect modification or biological interaction: two component causes can be said to interact synergistically simply if there is one sufficient cause that contains both of them. There is full synergism if they are only involved in producing the disease through the sufficient cause in which they are both present, i.e., there is no other sufficient cause in which one of them is present rather than the one in which they interact (complete synergism between CC_2 and CC_3 but partial synergism between CC_4 and CC_5). One of the interesting features of this model is that it clearly distinguishes the concept of interaction – a biological concept which can be represented under this framework – from the concept of confounding – a phenomenon which is introduced by the observer and has no biological role in disease causation.

Although the sufficient and component causes model has been used mainly as a conceptual framework for causal thinking, namely for teaching purposes, an interesting application to data from the European Prospective Investigation into Cancer and Nutrition has been published (Hoffmann et al 2006). The authors identified all possible combinations in a set of known component causes of myocardial infarction (smoking, hypertension, obesity and lack of exercise). Every possible combination was considered to be part of a set of sufficient causes and the population attributable fraction regarding that combination of factors was taken as a measure of the proportion of disease attributable to that class of sufficient causes in the population. The authors argue that beyond its theoretical contribution, by allowing for the modeling of sufficient causes without necessarily knowing all of the component causes, the model may be used to guide public health interventions.

2.3 Counterfactual model

According to counterfactual theories in philosophy, causation may be reflected upon by hypothesizing what would have occurred had the conditions been different from the actual conditions observed. In terms of counterfactual conditionals, the meaning of a cause A can be defined in the form "If A had not occurred, C would not have occurred" (Menzies 2009). Although it had been long explored in the philosophical literature, the application of the counterfactual model, or potential outcomes model, to epidemiologic research is recent (Greenland & Robins 1986).

According to the counterfactual approach, when assessing whether an exposure causes an outcome, most of the time we are interested, even if not explicitly stated, in comparing the occurrence of the outcome when the exposure is present with its occurrence if the exposure was absent and all other factors remained equal (Maldonado & Greenland 2002). If we could compare these outcomes and they were different, we would conclude that there was a causal relation between the exposure and the outcome. In other words, the counterfactual ideal contrasts the occurrence of the outcome in the actual exposure status (the observed, **factual** outcome) with the occurrence of outcome in the **same** individual or population had the exposure not been present (the unobserved, **counterfactual** outcome).The counterfactual or potential outcomes model formalizes this contrast in terms of epidemiologic research.

At the individual level, subjects may be classified according to four susceptibility types under the model, defined according to the combination of both potential outcomes (factual and counterfactual). Even though we can only observe factual outcomes, each individual can theoretically be classified according to two responses: the occurrence of the outcome had he been exposed and the occurrence of the outcome had he not been exposed. Table 1 shows the occurrence of the outcome according to each susceptibility type. If we consider an adverse outcome, susceptibility type 1 designates individuals who will develop the outcome whether or not they are exposed – they are **doomed**. Type 2 groups individuals who will develop the outcome if they are exposed but not if they are unexposed - **causative exposure**; among these, the exposure has a causal effect on the outcome. Individuals in whom the outcome will occur if the exposure is absent but not if the exposure is present are grouped in type 3 - **preventive exposure**. Finally, subjects are called **immune** (type 4) if they will not develop the outcome during the observation period whether or not they are exposed. It should then be noted that in types 1 and 4 factual and counterfactual outcomes are the same, i.e., the exposure has no effect on the occurrence of the outcome. The presence of the exposure relative to its absence affects only types 2 and 3.

Counterfactual susceptibility type	Occurrence of outcome if exposed	Occurrence of outcome if not exposed
Type 1 – doomed	1	1
Type 2 – exposure causal	1	0
Type 3 – exposure preventive	0	1
Type 4 – immune	0	0

Table 1. Counterfactual susceptibility types classified according to the occurrence of outcome in factual/counterfactual exposure conditions. Legend – 1: present, 0: absent

Although we can admit the theoretical existence of a counterfactual contrast, for each individual we can observe only one outcome while the other remains, by definition, unobserved. Therefore, we cannot classify an individual into his susceptibility type. Since potential outcomes include a response that would have been observed in an exposure experience that did not actually take place, if individual A is exposed and the outcome occurs we can classify him as type 1 or 2 but we will not be able to distinguish between these two types. Therefore, we will not be able to differentiate between an individual who would have developed the outcome whether or not he had been exposed and one who developed the outcome because of the exposure. This non-identifiability issue arises from the fact that the same observed association (occurrence of the outcome given exposure) may originate from both causal (type 2) and noncausal (type 1) relations. In the same way, if the outcome is not observed among an exposed individual B, that information does not allow for the distinction between an individual in whom the exposure was preventive and one who was immune to the outcome (types 3 and 4, respectively).

Greenland & Robins add that the observation that an exposed individual develops the outcome while an unexposed does not develop the outcome has no causal interpretation unless the assumption that these two individuals belong to the same susceptibility type is added to the observed data. This assumption implies that individuals are exchangeable, i.e. that if individual A had been unexposed and individual B had been exposed the same overall result would have been observed. Confounding becomes then equivalent to a lack of exchangeability and therefore may be defined in terms of comparability. Because counterfactual parameters are by definition unobservable, the exchangeability assumption is not verifiable. In practical terms this means that the comparability between exposed and unexposed subjects (i.e. the magnitude of confounding) cannot be measured directly from observed data.

Although this model is applicable to the individual, the unobserved nature of counterfactuals impairs causal inference at the individual level. Therefore, in epidemiology, we are frequently interested in estimating average effects in a population of an exposure in the occurrence of a disease. In a population, the counterfactual reasoning translates into comparing the probability of the occurrence of the outcome if the entire population had one exposure distribution (Figure 2, A1) with the probability of occurrence if the entire population had an alternative exposure distribution (A0). A causal effect would be present if these counterfactual probabilities were different (Maldonado & Greenland 2002).

It should be emphasized that under the counterfactual approach, an effect measure compares the frequency of the outcome under **two** exposure distributions, but in **one** target population during **one** etiologic time period. Since the true causal effect is by its counterfactual definition never observable (because the two exposure distributions cannot occur simultaneously in the same target population), causal inference has to rely on the contrast between the actual frequency of the outcome in the observed target population (A1) and the actual frequency of the outcome in a population which is a **substitute** for the counterfactual disease frequency in the target population with regard to the exposure under study (B0). The more similar the substitute is to the target population the more likely it is that the true causal contrast ($risk_{A1}/risk_{A0}$) is reflected in the estimated contrast ($risk_{A1}/risk_{B0}$). Bias in etiologic studies may be seen as resulting from the existence of relevant differences between the substitute and the target.

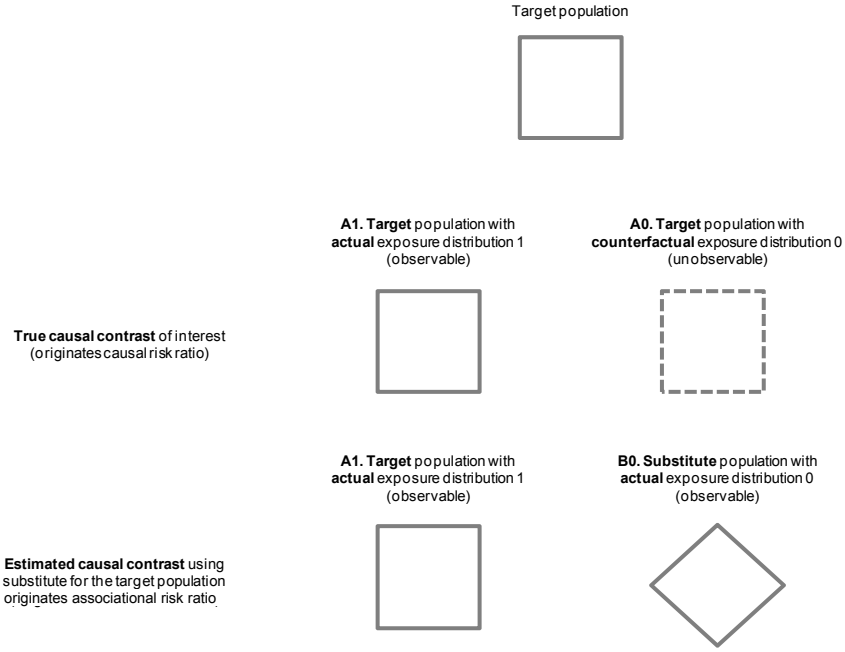


Fig. 2. Ideal causal (counterfactual) contrast and estimated causal contrast.

In the counterfactual framework, randomization can be seen as a means of obtaining the observed contrast as close as possible to the counterfactual ideal. If we assume perfect randomization and no random error, both groups (exposed to the index intervention and unexposed) are as similar as possible with regard to measured and unmeasured factors. As a consequence, the probability of developing the outcome among the control group equals the probability of developing the outcome in the intervention group had the latter not received the intervention (counterfactual). In this circumstance, groups are considered exchangeable and the causal risk ratio is accurately estimated by the associational risk ratio.

An application of the counterfactual framework to discuss selection criteria and generalizability of safety conclusions in randomized controlled trials has been recently published (Weisberg et al 2009). For a number of reasons, in trials of pharmacological interventions it is usual to exclude individuals with increased probability of drop-out or adverse events. This option is chosen at the expense of generalizability regarding the safety of the intervention, since selection probability is dependent on the baseline risk of adverse event. In their work, Weisberg *et al* argue that trials usually screen out individuals at high risk of an adverse event by measuring indicators of such risk. They make the point that those indicators are also predictive of the probability of experiencing adverse events in the **absence** of treatment (i.e. under placebo) since they are measured before the intervention starts. The authors assume that these indicators remain better predictors of the probability of adverse event under placebo than if treatment is introduced. As a result of those strict selection criteria, individuals most likely to have an adverse event if placed in the placebo arm of the trial have higher probability of being excluded. In counterfactual terms, this

means that there is an underrepresentation of individuals for whom the intervention would be preventive, i.e. who would have the adverse event if under placebo but not if they received the active treatment.

Table 2 presents the hypothetical results proposed by the authors for a trial where the outcome is an adverse event. Two scenarios are posed: one in which all counterfactual susceptibility types regarding the adverse event under study have the same selection probability, which corresponds to their prevalence in the target population, and another in which there is selective undersampling of individuals who would have had the event under placebo, i.e. doomed and preventive susceptibility types. In this example, the authors use counterfactual susceptibility types to illustrate that the ratio of the risks of adverse event between the two arms may be biased and a conclusion of a causal effect may be derived when the observed association is due only to selection criteria which originate a dependence between counterfactual susceptibility types and the probability of entering the trial.

Under non-differential selection criteria					
Counterfactual susceptibility type	Proportion in the population	Selection probability	Number of participants in each arm	Adverse events under active treatment	Adverse events under placebo
Doomed	6%	100%	15	15	15
Causal	4%	100%	10	10	0
Preventive	4%	100%	10	0	10
Immune	86%	100%	215	0	0
			Risk of adverse event	*10.0%*	*10.0%*
Under differential selection criteria					
Counterfactual susceptibility type	Proportion in the population	Selection probability	Number of participants in each arm	Adverse events under active treatment	Adverse events under placebo
Doomed	6%	20%	3	3	3
Causal	4%	80%	8	8	0
Preventive	4%	20%	2	0	2
Immune	86%	80%	172	0	0
			Risk of adverse event	*5.9%*	*2.7%*

Table 2. Selection probability for a hypothetical randomized trial according to counterfactual susceptibility types (adapted from Weisberg 2009).

This is an interesting example of the use of counterfactual thinking for illustrating selection bias. Although one may assume, contrarily to the authors, that individuals who are at greater risk of developing adverse events under treatment (doomed and causal rather that doomed and preventive) are underrepresented, the issue of different selection probabilities given different susceptibility types would remain relevant.

More often than not, experimental studies are not feasible and causal inference relies on observational data. In non-randomized studies, confounding emerges if the chosen substitute does not accurately represent the target population under the counterfactual

condition, i.e. if the two groups of contrasting exposure distributions that are compared with respect to the occurrence of the outcome are not exchangeable. Such inadequate substitute may originate relevant differences between the observed (associational) risk ratio and the true (causal) risk ratio.

Today, there is growing use of the counterfactual model of causation in epidemiologic research. The application of the counterfactual framework to epidemiologic research has been subject to discussion (Dawid 2002; Elwert & Winship 2002; Kaufman & Kaufman 2002; Shafer 2002). In practice, it should be noted that this framework does not aim at clarifying mechanisms of disease causation and that there is a potential for impossible counterfactuals, which are not amenable to intervention and have therefore limited interest in public health. Nevertheless, the model provides, for individuals or populations, a causal interpretation to measures of association. More importantly, the counterfactual ideal is relevant in the design and analysis of etiologic studies since it provides a framework for choosing the target population as well as an appropriate substitute for that target. It is also an important approach to clarify threats to validity.

2.4 Causal diagrams

One of the most intuitive ways of representing and communicating hypothetical causal relations between epidemiological variables is to visually depict the paths believed to relate them, namely exposure, outcome and possible confounders. For a long time the use of such visual tools to assist causal inference was informal and lacked theoretical support. The development of computer science and the corresponding need for improvement of the quality of the decision process on the basis of the relations between previously defined and empirical data from complex systems led to the development of a solid theoretical basis for the use of graphical models (Pearl 1995).

This body of work allowed for the formal application of causal diagrams outside the artificial intelligence domain to a number of areas where causal inference relies on the combination of causal assumptions with observational data. Causal diagrams have been increasingly used in epidemiology and have proved to be a helpful approach to conceptualize research questions and analytical issues (Greenland et al 1999). Indeed, one of the most interesting features of these diagrams is that, by depicting qualitative and nonparametric assumptions of causal mechanisms linking variables in a dataset and knowing a number of mathematical rules, it is possible to characterize the nature of common systematic errors in causal inference, as well as to guide study design.

Causal diagrams used in epidemiology are known as **directed acyclic graphs** (DAGs). The relations between variables in a DAG are translated in a set of formal rules known as ***d*-separation rules** (where *d* means "directional"), which are used to judge whether variables are associated (*d*-connected) or independent (*d*-separated) (Pearl 2009). In a DAG, variables are named **nodes** and are linked by arrows called **edges**. Since a variable cannot be, at the same instant, a cause and a consequence of another variable there are no cycles in DAGs. In the following example, E and O are nodes connected by an edge. In DAG terminology E is a parent or ancestor of O, and O is a child or descendant of E.

$$E \rightarrow O$$

A **path** is a sequence of edges that connect two variables regardless of the direction of the edges, such as the following path that links variables A and O. This path is called **unblocked** because there are no head to head arrows colliding along the path:

$$A \rightarrow X \rightarrow E \rightarrow O$$

If two edges meet head to head the node where they meet is called a **collider**. In the following example, where Y is a collider, the path between A and Y is unblocked (A and Y are *d*-connected) but the path between A and O is blocked by Y (A and O are *d*-separated):

$$A \rightarrow X \rightarrow E \rightarrow Y \leftarrow O$$

Any one of these nodes (or variables) may be **conditioned on** or, in other words, it may be set at a specific value. In classic epidemiologic language, this is most of the times equivalent to adjustment for or stratification according to X. If a **non-collider** such as X is conditioned on, as shown by the square symbol around it, the path between A and Y becomes **blocked** and A and Y become *d*-separated:

$$A \rightarrow \boxed{X} \rightarrow E \rightarrow Y \leftarrow O$$

If, however, a **collider** such as Y is conditioned on, the path between A and O becomes **unblocked** and A and O become *d*-connected:

$$A \rightarrow X \rightarrow E \rightarrow \boxed{Y} \leftarrow O$$

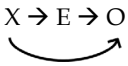
Pearl provided also a very clear example of the meaning of conditioning on a collider. Suppose that there are two reasons for a car not starting: not having fuel and having a dead battery, according to the following diagram:

$$\text{Dead battery} \rightarrow \text{Car does not start} \leftarrow \text{No fuel}$$

These causes are marginally independent, i.e., having information on one of them tells us nothing about the other (they are *d*-separated). Indeed, knowing that the battery is dead does not improve our prediction of whether or not there is fuel. However, if the collider is conditioned on, i.e., if we known that the car does not start, knowing that the battery is not dead tells us that there must be no fuel. Dead battery and no fuel, which were marginally independent, become dependent (*d*-connected) by conditioning on their common effect. If the descendant of a collider such as Z is conditioned on, the path between A and O also becomes unblocked:

$$\begin{array}{c} A \rightarrow X \rightarrow E \rightarrow Y \leftarrow O \\ \qquad\qquad\quad \downarrow \\ \qquad\qquad\quad \boxed{Z} \end{array}$$

A DAG is causal only if it includes the common causes of any two nodes represented (Markov condition). In the previous case, if there is a variable X which is a common cause of E and O it must be depicted in the diagram. This condition simplifies the diagram since it implies that any variable which is not a common cause of two or more variables in the DAG does not need to be depicted in the graph.



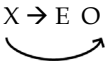
In the previous diagram the path that connects E to O through X (E←X→O) is called a **backdoor path**. If the common cause of E and O is conditioned on, that backdoor path will be closed. Consequently, if X is the only common cause of E and O, after adjustment for X, the statistical association found between E and O will be a result of the true causal effect of E on O, which is equivalent to eliminating confounding.

Using the above-mentioned rules, and assuming no random error, statistical associations between variables can be found in the three following causal diagram scenarios:

1. They are cause (E) and effect (O):

   E→O

2. They share common causes (X) that have not been conditioned on (open backdoor path):



3. They share common effects (Y) that have been conditioned on:

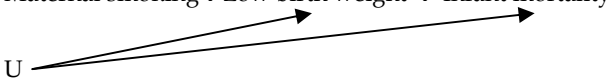
   E→Y←O

While the first alternative depicts a **true causal effect** of E on O, the second scenario is the graphical representation of **confounding** and the third translates **selection bias**.

A very interesting example of the application of causal diagrams to judge the appropriateness of conditioning on available variables has been presented to explain the so-called "birth weight paradox" (Hernandez-Diaz et al 2006). In perinatal epidemiology, birth weight is a routinely collected variable because of its ability to predict adverse outcomes during infancy. Probably as a result of its importance and wide availability, birth weight is frequently an adjustment or stratification variable in the analysis of the effects of several other exposures on infant mortality.

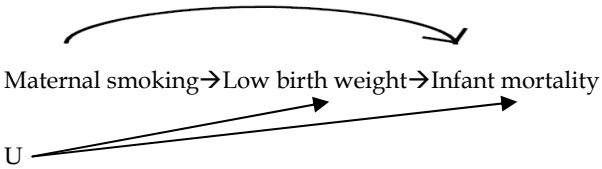
From the comparison of the associations found among low birth weight infants with those found in normal weight infants emerged what was called the birth weight paradox: the associations between known adverse exposures such as maternal smoking and infant mortality was weaker among low birth weight infants than among normal weight children. In other words, although it was known that smoking caused low birth weight and increased infant mortality, it was estimated that the effect of smoking on mortality was greater in normal weight infants. As a consequence, it was hypothesized that maternal smoking could be protective against infant mortality in low birth weight children.

To explain this apparent paradox, Hernandez-Diaz *et al* proposed several DAGs depicting *a priori* assumptions about the causal relations between variables. They proposed that even using a simple set of causal assumptions it may be shown that bias may arise from adjustment to birth weight. Convincing evidence exists that there are common causes of low birth weight and mortality (U), such as birth defects and malnutrition. In those circumstances, the most simplistic DAG that could be drawn would be the following:



Low birth weight is a collider in this graph, since it is a common effect of smoking as well as of other causes (U). Conditioning on birth weight will create a spurious dependence between maternal smoking and the other low birth causes. Indeed, if we know that a child had low birth weight and that the mother never smoked, then it is more likely that the child had low birth weight because of another (probably more serious) cause than maternal smoking, such as a birth defect. That other cause will probably be a stronger determinant of infant mortality than smoking.

Therefore, in low birth weight infants, smoking may appear protective of infant mortality only because, after birth weight stratification, it indicates lower probability of other causes of low birth weight and therefore of infant mortality. This bias will emerge even if the "true" causal DAG includes an effect of low birth weight on infant mortality and/or an effect of maternal smoking on mortality not mediated by birth weight, as shown in the following diagram:



This interesting example of the introduction of selection bias in the analysis of data illustrates the usefulness of causal diagrams in distinguishing between variables that measure common causes and those which measure common effects. In fact, one of the points made by the authors of the example referred above is that the choice of variables to include in regression modeling should not be driven by their availability or biological relevance alone, but after a hypothesis has been formulated on what their role in the causal pathway might be.

Another important example of a decision in data analysis where causal diagrams may be of importance is the analysis of change, which is of great importance with the increasing abundance of prospectively collected data. Imagine we are studying the physiological increase in bone strength (measured through bone mineral density) throughout adolescence. We are interested in assessing whether adiposity in early adolescence (measured as body fat mass) has an effect on the extent of bone strength increase between early and late adolescence. Suppose that we have a dataset with the following variables:

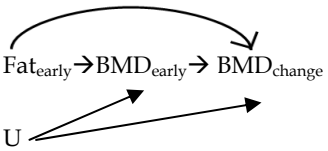
- Bone mineral density in early adolescence (BMD_{early})
- Bone mineral density increase from early to late adolescence (BMD_{change})
- Total body fat mass in early adolescence (Fat_{early})

There is a body of evidence indicating that total body fat has an overall positive effect on bone strength at the same age. Additionally, there is also evidence that baseline bone density has an effect on the change in bone strength throughout the following years. Our causal DAG could be the following:

$Fat_{early} \rightarrow BMD_{early} \rightarrow BMD_{change}$

What we want to know is if there might be an effect of fat on BMD change that is independent of bone mineral density in early adolescence, as depicted in the following diagram:

One of the questions we frequently have when facing this kind of research problems is whether or not we should adjust for baseline characteristics, in this case for BMD in early adolescence. Using different examples, the point has been made that it is very likely that there are unmeasured common causes of variables shown in the previous diagram (Glymour et al 2005; VanderWeele 2009). According to rules previously presented, any common causes of two variables in a DAG have to be depicted. That means that the previous DAG is probably not a good representation of the true causal structure, and that the following diagram could be more consistent with previous expert knowledge:



This addition of unmeasured common causes to the diagram turns BMD_{early} into a collider, which means that adjusting for (or conditioning on) that variable creates a spurious association between those common causes and adiposity, which were marginally independent before adjustment. This means that an adjusted estimate may be biased, since we may be in truth measuring the association between those unmeasured common causes and bone strength change and estimating it as the true effect of adiposity on bone strength change.

If we consider that one of those unmeasured common causes may be lean body mass, we may see conditioning on a collider similarly to the car example: if an adolescent has low body fat and high bone strength, that may be due to increase lean mass, which has documented positive effect on bone properties. Adjusting for baseline BMD (a collider) could therefore originate biased effect estimates. Only by identifying and measuring all common causes of any two variables in the diagram would we be able to adjust for them, blocking all backdoor paths and thus eliminating confounding.

As seen from the examples above, by applying a small set of formal mathematical rules these diagrams provide a tool for identifying sources of bias in study design or in the analysis of results, thus providing an interesting framework for causal inference. DAGs have the additional advantage of clearly depicting assumptions about the causal structure of data, thus improving clarity in the communication of hypothesized causal relations between variables. Since they are nonparametric qualitative models which do not by themselves provide information on the magnitude of effects, DAGs may be used to assess to which extent previous assumptions are compatible with observed data.

Causal diagrams should not be expected to provide the answer to causal questions. DAGs are useful exactly because we will most probably never know the "true" DAG for most causal mechanisms. These diagrams provide a common framework for clarifying causal assumptions and guiding study design and data analysis in such a way that it is not contradictory with those assumptions.

3. Conclusion

Epidemiologic research is primarily driven by the observation of difference, which in human populations virtually never corresponds to an ideal contrast. Epidemiologists then try to explain that difference and to identify the factors that can be acted upon to improve population health.

The development and widespread use of statistical modeling techniques has dramatically improved our ability to efficiently quantify statistical associations between variables. It has allowed us to model relations between enormous numbers of variables. In the context of etiologic research, the magnitude and even the statistical significance of the associations estimated have been used as evidence for causation, while a more formal causality assessment has frequently gone undiscussed.

In order to quantify the role of different types of findings in the appraisal of causality, an interesting experience conducted among 159 epidemiologists, where each subject was shown computer-simulated summaries of evidence of the relation between an exposure and an outcome (Holman et al 2001). In this study, the factors with the strongest influence in causal attribution by epidemiologists were statistical significance and refutation of alternative explanations, followed by strength of association and coherence.

The frameworks presented strengthen the point that etiologic research aims at disclosing causal relations rather than co-occurring characteristics. Assumptions regarding causation are present in virtually all domains of scientific research. In areas where controlled experiments are admissible, the observation of difference may frequently be taken as evidence for causation and the theory underlying causal thinking remains implicit. However, causal inference in epidemiology can seldom be a result of such ideal experiments. The observational nature of most epidemiologic research is probably the main reason for the search for frameworks to guide causal inference in the study of the etiology of human disease.

None of the models presented can be expected to uncover the complexity of disease causation. Nevertheless, these frameworks are important contributions for the design and analysis of epidemiological studies, as well as for integrating observed data and prior knowledge with the purpose of judging whether or not a true causal effect is the best explanation for differences observed.

4. Acknowledgment

Project grant FCT-PTDC/SAU-ESA/108407/2008 and individual grant SFRH/BD/40656/2007 from the Portuguese Foundation for Science and Technology are gratefully acknowledged.

5. References

Bosch FX, Lorincz A, Munoz N, Meijer CJ, Shah KV. (2002). The causal relation between human papillomavirus and cervical cancer. J Clin Pathol 55:244-65

Broadbent A. (2009). Causation and models of disease in epidemiology. Stud Hist Philos Biol Biomed Sci 40:302-11

Dawid AP. (2002). Counterfactuals: help or hindrance? Int J Epidemiol 31:429-30; discussion 35-8

Dictionary.com (2011). cause. (n.d.). Dictionary.com Unabridged. Retrieved September 23, 2011, from Dictionary.com website: http://dictionary.reference.com/browse/cause.

Elwert F, Winship C. (2002). Population versus individual level causal effects. Int J Epidemiol 31:432-4; discussion 5-8

Glymour MM, Weuve J, Berkman LF, Kawachi I, Robins JM. (2005). When is baseline adjustment useful in analyses of change? An example with education and cognitive change. Am J Epidemiol 162:267-78

Greenland S, Pearl J, Robins JM. (1999). Causal diagrams for epidemiologic research. Epidemiology 10:37-48

Greenland S, Robins JM. (1986). Identifiability, exchangeability, and epidemiological confounding. Int J Epidemiol 15:413-9

Hernandez-Diaz S, Schisterman EF, Hernan MA. (2006). The birth weight "paradox" uncovered? Am J Epidemiol 164:1115-20

Hill AB. (1965). The environment and disease: association or causation? Proc R Soc Med 58:295-300

Hoffmann K, Heidemann C, Weikert C, Schulze MB, Boeing H. (2006). Estimating the proportion of disease due to classes of sufficient causes. Am J Epidemiol 163:76-83

Holman CD, Arnold-Reed DE, de Klerk N, McComb C, English DR. (2001). A psychometric experiment in causal inference to estimate evidential weights used by epidemiologists. Epidemiology 12:246-55

Kaufman JS, Kaufman S. (2002). Estimating causal effects. Int J Epidemiol 31:431-2; discussion 5-8

Krieger N. (1994). Epidemiology and the web of causation: has anyone seen the spider? Soc Sci Med 39:887-903

Luedeking R. (1884). The Chief Local Factors in the Causation of Disease and Death. Public Health Pap Rep 10:329-31

Maldonado G, Greenland S. 2002. Estimating causal effects. Int J Epidemiol 31:422-9

Menzies P. (2009). "Counterfactual Theories of Causation", The Stanford Encyclopedia of Philosophy (Fall 2009 Edition), Edward N. Zalta (ed.), URL = <http://plato.stanford.edu/archives/fall2009/entries/causation-counterfactual/>

Parascandola M, Weed DL. (2001). Causation in epidemiology. J Epidemiol Community Health 55:905-12

Pearl J. (1995). Causal diagrams for empirical research. Biometrika 82:669-710

Pearl J. (2009). Causality: Models, Reasoning and Inference: Cambridge University Press; 2 edition.

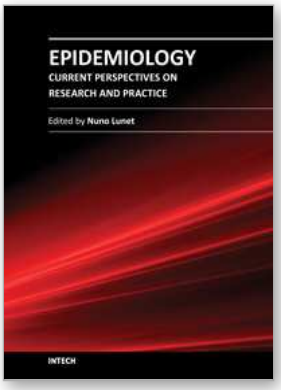
Rothman K, Greenland S, Lash T. (2008). Modern Epidemiology: Lippincott Williams & Wilkins. Philadelphia, 2008.

Rothman KJ. (1976). Causes. Am J Epidemiol 104:587-92

Shafer G. (2002). Estimating causal effects. Int J Epidemiol 31:434-5; discussion 5-8

VanderWeele TJ. (2009). Marginal structural models for the estimation of direct and indirect effects. Epidemiology 20:18-26.

Weisberg HI, Hayden VC, Pontes VP. (2009). Selection criteria and generalizability within the counterfactual framework: explaining the paradox of antidepressant-induced suicidality? Clin Trials 6:109-18.



Epidemiology - Current Perspectives on Research and Practice
Edited by Prof. Nuno Lunet

ISBN 978-953-51-0382-0
Hard cover, 208 pages
Publisher InTech
Published online 13, March, 2012
Published in print edition March, 2012

This special issue resulted from the invitation made to selected authors to contribute with an overview of a specific subject of their choice, and is based on a collection of papers chosen to exemplify some of the interests, uses and views of the epidemiology across different areas of research and practice. Rather than the comprehensiveness and coherence of a conventional textbook, readers will find a set of independent chapters, each of them of a great interest in their own specialized areas within epidemiology. Taken together, they illustrate the contrast between the attempt to extend the limits of applicability of epidemiological research, and the "regular" scientific activity in this field or an applied epidemiology. Epidemiologists with different levels of expertise and interests will be able to find informative and inspiring readings among the chapters of this book.

INTECH
open science | open minds

InTech Europe
University Campus STeP Ri
Slavka Krautzeka 83/A
51000 Rijeka, Croatia
Phone: +385 (51) 770 447
Fax: +385 (51) 686 166
www.intechopen.com

InTech China
Unit 405, Office Block, Hotel Equatorial Shanghai
No.65, Yan An Road (West), Shanghai, 200040, China
中国上海市延安西路65号上海国际贵都大饭店办公楼405单元
Phone: +86-21-62489820
Fax: +86-21-62489821