CONTINUING PROFESSIONAL EDUCATION

A definition of causal effect for epidemiological research

M A Hernán

J Epidemiol Community Health 2004;**58**:265–271. doi: 10.1136/jech.2002.006361

Estimating the causal effect of some exposure on some outcome is the goal of many epidemiological studies. This article reviews a formal definition of causal effect for such studies. For simplicity, the main description is restricted to dichotomous variables and assumes that no random error attributable to sampling variability exists. The appendix provides a discussion of sampling variability and a generalisation of this causal theory. The difference between association and causation is described—the redundant expression ''causal effect'' is used throughout the article to avoid confusion with a common use of ''effect'' meaning simply statistical association—and shows why, in theory, randomisation allows the estimation of causal effects without further assumptions. The article concludes with a discussion on the limitations of randomised studies. These limitations are the reason why methods for causal inference from observational data are needed.

Correspondence to:
Dr M Hernán, Department of Epidemiology, Harvard School of Public Health, 677 Huntington Avenue, Boston, MA 02115, USA; miguel_hernan@post.harvard.edu

Accepted for publication
29 August 2003

INDIVIDUAL CAUSAL EFFECTS

Zeus is a patient waiting for a heart transplant. On 1 January, he received a new heart. Five days later, he died. Imagine that we can somehow know, perhaps by divine revelation, that had Zeus not received a heart transplant on 1 January (all others things in his life being unchanged) then he would have been alive five days later. Most people equipped with this information would agree that the transplant caused Zeus' death. The intervention had a causal effect on Zeus' five day survival.

Another patient, Hera, received a heart transplant on 1 January. Five days later she was alive. Again, imagine we can somehow know that had Hera not received the heart on 1 January (all other things being equal) then she would still have been alive five days later. The transplant did not have a causal effect on Hera's five day survival.

These two vignettes illustrate how human reasoning for causal inference works: we compare (often only mentally) the outcome when action A is present with the outcome when action A is absent, all other things being equal. If the two outcomes differ, we say that the action A has a causal effect, causative or preventive, on the outcome. Otherwise, we say that the action A has no causal effect on the outcome. In epidemiology, A is commonly referred to as exposure or treatment.

The next step is to make this causal intuition of ours amenable to mathematical and statistical analysis by introducing some notation. Consider a dichotomous exposure variable A (1: exposed, 0: unexposed) and a dichotomous outcome variable Y (1: death, 0: survival). Table 1 shows the data from a heart transplant observational study with 20 participants. Let $Y_{a=1}$ be the outcome variable that would have been observed under the exposure value $a = 1$, and $Y_{a=0}$ the outcome variable that would have been observed under the exposure value $a = 0$. (Lowercase a represents a particular value of the variable A.) As shown in table 2, Zeus has $Y_{a=1} = 1$ and $Y_{a=0} = 0$ because he died when exposed but would have survived if unexposed.

We are now ready to provide a formal definition of causal effect for each person: exposure has a causal effect if $Y_{a=0} \neq Y_{a=1}$. Table 2 is all we need to decide that the exposure has an effect on Zeus' outcome because $Y_{a=0} \neq Y_{a=1}$, but not on Hera's outcome because $Y_{a=0} = Y_{a=1}$. When the exposure has no causal effect for any subject—that is, $Y_{a=0} = Y_{a=1}$ for all subjects—we say that the *sharp causal null hypothesis* is true.

The variables $Y_{a=1}$ and $Y_{a=0}$ are known as potential outcomes because one of them describes the subject's outcome value that would have been observed under a potential exposure value that the subject did not actually experience. For example, $Y_{a=0}$ is a potential outcome for exposed Zeus, and $Y_{a=1}$ is a potential outcome for unexposed Hera. Because these outcomes would have been observed in situations that did not actually happen (that is, in counter to the fact situations), they are also known as *counterfactual outcomes*. For each subject, one of the counterfactual outcomes is actually factual—the one that corresponds to the exposure level or treatment regimen that the subject actually received. For example, if $A = 1$ for Zeus, then $Y_{a=1} = Y_{a=A} = Y$ for him.

The fundamental problem of causal inference should now be clear. Individual causal effects are defined as a contrast of the values of counterfactual outcomes, but only one of those values is observed. Table 3 shows the observed data and each subject's observed counterfactual outcome: the one corresponding to the exposure value actually experienced by the subject. All other counterfactual outcomes are missing. The unhappy conclusion is that, in general, individual causal effects cannot be identified because of missing data.

POPULATION CAUSAL EFFECT

We define the probability $\Pr[Y_a = 1]$ as the proportion of subjects that would have developed

Table 1 Data from a study with dichotomous exposure *A* and outcome *Y*

ID	*A*	*Y*
Rheia	0	0
Kronos	0	1
Demeter	0	0
Hades	0	0
Hestia	1	0
Poseidon	1	0
Hera	1	0
Zeus	1	1
Artemis	0	1
Apollo	0	1
Circe	0	0
Ares	1	1
Athene	1	1
Eros	1	1
Aphrodite	1	1
Prometheus	1	1
Selene	1	1
Hermes	1	0
Eos	1	0
Helios	1	0

Table 2 Counterfactual outcomes of subjects in a study with dichotomous exposure *A* and outcome *Y*

ID	$Y_{a=0}$	$Y_{a=1}$
Rheia	0	1
Kronos	1	0
Demeter	0	0
Hades	0	0
Hestia	0	0
Poseidon	1	0
Hera	0	0
Zeus	0	1
Artemis	1	1
Apollo	1	0
Circe	0	1
Ares	1	1
Athene	1	1
Eros	0	1
Aphrodite	0	1
Prometheus	0	1
Selene	1	1
Hermes	1	0
Eos	1	0
Helios	1	0

the outcome Y had all subjects in the population of interest received exposure value a. We also refer to $\Pr[Y_a = 1]$ as the risk of Y_a. The exposure has a causal effect in the population if $\Pr[Y_{a=1} = 1] \neq \Pr[Y_{a=0} = 1]$.

Suppose that our population is comprised by the subjects in table 2. Then $\Pr[Y_{a=1} = 1] = 10/20 = 0.5$, and $\Pr[Y_{a=0} = 1] = 10/20 = 0.5$. That is, 50% of the patients would have died had everybody received a heart transplant, and 50% would have died had nobody received a heart transplant. The exposure has no effect on the outcome at the population level. When the exposure has no causal effect in the population, we say that the *causal null hypothesis* is true.

Unlike individual causal effects, population causal effects *can* sometimes be computed—or, more rigorously, consistently estimated (see appendix)—as discussed below. Hereafter we refer to the "population causal effect" simply as "causal effect". Some equivalent definitions of causal effect are

(a) $\Pr[Y_{a=1} = 1] - \Pr[Y_{a=0} = 1] \neq 0$

(b) $\Pr[Y_{a=1} = 1] / \Pr[Y_{a=0} = 1] \neq 1$

(c) $(\Pr[Y_{a=1} = 1] / \Pr[Y_{a=1} = 0]) / (\Pr[Y_{a=0} = 1] / \Pr[Y_{a=0} = 0]) \neq 1$

where the left hand side of inequalities (a), (b), and (c) is the causal risk difference, risk ratio, and odds ratio, respectively. The causal risk difference, risk ratio, and odds ratio (and other causal parameters) can also be used to quantify the strength of the causal effect when it exists. They measure the same causal effect in different scales, and we refer to them as *effect measures*.

ASSOCIATION AND CAUSATION

To characterise association, we first define the probability $\Pr[Y = 1 | A = a]$ as the proportion of subjects that developed the outcome Y among those subjects in the population of interest that happened to receive exposure value a. We also refer to $\Pr[Y = 1 | A = a]$ as the risk of Y given $A = a$. Exposure and outcome are associated if $\Pr[Y = 1 | A = 1] \neq \Pr[Y = 1 | A = 0]$. In our population of

Table 3 Data and observed counterfactual outcomes from a study with dichotomous exposure *A* and outcome *Y*

ID	*A*	*Y*	$Y_{a=0}$	$Y_{a=1}$
Rheia	0	0	0	?
Kronos	0	1	1	?
Demeter	0	0	0	?
Hades	0	0	0	?
Hestia	1	0	?	0
Poseidon	1	0	?	0
Hera	1	0	?	0
Zeus	1	1	?	1
Artemis	0	1	1	?
Apollo	0	1	1	?
Circe	0	0	0	?
Ares	1	1	?	1
Athene	1	1	?	1
Eros	1	1	?	1
Aphrodite	1	1	?	1
Prometheus	1	1	?	1
Selene	1	1	?	1
Hermes	1	0	?	0
Eos	1	0	?	0
Helios	1	0	?	0

table 1, exposure and outcome are associated because $\Pr[Y=1|A=1]=7/13$, and $\Pr[Y=1|A=0]=3/7$. Some equivalent definitions of association are

(a) $\Pr[Y=1|A=1]-\Pr[Y=1|A=0]\neq 0$

(b) $\Pr[Y=1|A=1]/\Pr[Y=1|A=0]\neq 1$

(c) $(\Pr[Y=1|A=1]/\Pr[Y=0|A=1])/(\Pr[Y=1|A=0]/\Pr[Y=0|A=0])\neq 1$

where the left hand side of the inequalities (a), (b), and (c) is the associational risk difference, risk ratio, and odds ratio, respectively. The associational risk difference, risk ratio, and odds ratio (and other association parameters) can also be used to quantify the strength of the association when it exists. They measure the same association in different scales, and we refer to them as *association measures*.

When A and Y are not associated, we say that A does not predict Y, or vice versa. Lack of association is represented by $Y \coprod A$ (or, equivalently, $A \coprod Y$), which is read as Y and A are independent.

Note that the risk $\Pr[Y=1|A=a]$ is computed using the subset of subjects of the population that meet the condition "having actually received exposure a" (that is, it is a conditional probability), whereas the risk $\Pr[Y_a=1]$ is computed using *all* subjects of the population had they received the counterfactual exposure a (that is, it is an unconditional or marginal probability). Therefore, association is defined by a different risk in two disjoint subsets of the population determined by the subjects' actual exposure value, whereas causation is defined by a different risk in the same subset (for example, the entire population) under two potential exposure values (fig 1). This radically different definition accounts for the well known adage "association is not causation." When an association measure differs from the corresponding effect measure, we say that there is *bias* or *confounding*.

COMPUTATION OF CAUSAL EFFECTS VIA RANDOMISATION

Unlike association measures, effect measures cannot be directly computed because of missing data (see table 3). However, effect measures can be computed—or, more rigorously, consistently estimated (see appendix)—in randomised experiments.

Suppose we have a (near-infinite) population and that we flip a coin for each subject in such population. We assign the subject to group 1 if the coin turns tails, and to group 2 if it turns heads. Next we administer the treatment or exposure of interest ($A=1$) to subjects in group 1 and placebo ($A=0$) to those in group 2. Five days later, at the end of the study, we compute the mortality risks in each group, $\Pr[Y=1|A=1]$ and $\Pr[Y=1|A=0]$. For now, let us assume that this randomised experiment is ideal in all other respects (no loss to follow up, full compliance with assigned treatment, blind assignment).

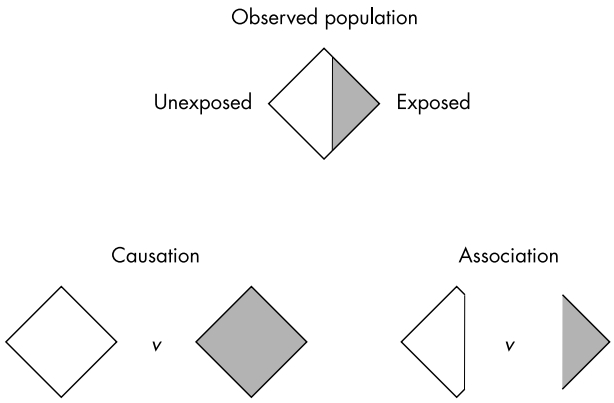

Figure 1 Causation is defined by a different risk in the entire population under two potential exposure values; association is defined by a different risk in the subsets of the population determined by the subjects' actual exposure value.

We will show that, in such a study, the observed risk $\Pr[Y=1|A=a]$ is equal to the counterfactual risk $\Pr[Y_a=1]$, and therefore the associational risk ratio equals the causal risk ratio.

First note that, when subjects are randomly assigned to groups 1 and 2, the proportion of deaths among the exposed, $\Pr[Y=1|A=1]$, will be the same whether subjects in group 1 receive the exposure and subjects in group 2 receive placebo, or vice versa. Because group membership is randomised, both groups are "comparable": which particular group got the exposure is irrelevant for the value of $\Pr[Y=1|A=1]$. (The same reasoning applies to $\Pr[Y=1|A=0]$.) Formally, we say that both groups are exchangeable.

Exchangeability means that the risk of death in group 1 would have been the same as the risk of death in group 2 had subjects in group 1 received the exposure given to those in group 2. That is, the risk under the potential exposure value a among the exposed, $\Pr[Y_a=1|A=1]$, equals the risk under the potential exposure value a among the unexposed, $\Pr[Y_a=1|A=0]$, for $a=0$ and $a=1$. An obvious consequence of these (conditional) risks being equal in all subsets defined by exposure status in the population is that they must be equal to the (marginal) risk under exposure value a in the whole population: $\Pr[Y_a=1|A=1]=\Pr[Y_a=1|A=0]=\Pr[Y_a=1]$. In other words, under exchangeability, the actual exposure does not predict the counterfactual outcome; they are independent, or $Y_a \coprod A$ for all values a. Randomisation produces exchangeability.

We are only one step short of showing that the observed risk $\Pr[Y=1|A=a]$ equals the counterfactual risk $\Pr[Y_a=1]$ in ideal randomised experiments. By definition, the value of the counterfactual outcome Y_a for subjects who actually received exposure value a is their observed outcome value Y. Then, among those who actually received exposure value a, the risk under the potential exposure value a is trivially equal to the observed risk. That is, $\Pr[Y_a=1|A=a]=\Pr[Y=1|A=a]$.

Let us now combine the results from the two previous paragraphs. Under exchangeability, $Y_a \coprod A$ for all a, the conditional risk among those exposed to a is equal to the marginal risk had the whole population been exposed to a: $\Pr[Y_a=1|A=1]=\Pr[Y_a=1|A=0]=\Pr[Y_a=1]$. And by definition of counterfactual outcome $\Pr[Y_a=1|A=a]=\Pr[Y=1|A=a]$. Therefore, the observed risk $\Pr[Y=1|A=a]$ equals the counterfactual risk $\Pr[Y_a=1]$. In ideal randomised experiments, association *is* causation. On the other hand, in non-randomised (for example, observational) studies association is not necessarily causation because of potential lack of exchangeability of exposed and unexposed subjects. For example, in our heart transplant study, the risk of death under no treatment is different for the exposed and the unexposed: $\Pr[Y_{a=0}=1|A=1]=7/13\neq\Pr[Y_{a=0}=1|A=0]=3/7$. We say that the exposed had a worse prognosis, and therefore a greater risk of death, than the unexposed, or that $Y_a \coprod A$ does not hold for $a=0$.

INTERVENTIONS AND CAUSAL QUESTIONS

We have so far assumed that the counterfactual outcomes Y_a exist and are well defined. However, that is not always the case.

Suppose women ($S=1$) have a greater risk of certain disease Y than men ($S=0$)—that is, $\Pr[Y=1|S=1]>\Pr[Y=1|S=0]$. Does sex S has a causal effect on the risk

of Y—that is, $\Pr[Y_{s=1}=1] > \Pr[Y_{s=0}=1]$? This question is quite vague because it is unclear what we mean by the risk of Y had everybody been a woman (or a man). Do we mean the risk of Y had everybody "carried a pair of X chromosomes", "been brought up as a woman", "had female genitalia", or "had high levels of oestrogens between adolescence and menopausal age"? Each of these definitions of the exposure "female sex" would lead to a different causal effect.

To give an unambiguous meaning to a causal question, we need to be able to describe the interventions that would allow us to compute the causal effect in an ideal randomised experiment. For example, "administer 30 μg/day of ethinyl estradiol from age 14 to age 45" compared with "administer placebo." That some interventions sound technically unfeasible or plainly crazy simply indicates that the formulation of certain causal questions (for example, the effect of sex, high serum LDL-cholesterol, or high HIV viral load on the risk of certain disease) is not always straightforward. A counterfactual approach to causal inference highlights the imprecision of ambiguous causal questions, and the need for a common understanding of the interventions involved.

LIMITATIONS OF RANDOMISED EXPERIMENTS

We now review some common methodological problems that may lead to bias in randomised experiments. To fix ideas, suppose we are interested in the causal effect of a heart transplant on one year survival. We start with a (near-infinite) population of potential recipients of a transplant, randomly allocate each subject in the population to either transplant ($A=1$) or medical treatment ($A=0$), and ascertain how many subjects die within the next year ($Y=1$) in each group. We then try to measure the effect of heart transplant on survival by computing the associational risk ratio $\Pr[Y=1|A=1]/\Pr[Y=1|A=0]$, which is theoretically equal to the causal risk ratio $\Pr[Y_{a=1}=1]/\Pr[Y_{a=0}=1]$. Consider the following problems:

- *Loss to follow up*. Subjects may be lost to follow up or drop out of the study before their outcome is ascertained. When this happens, the risk $\Pr[Y=1|A=a]$ cannot be computed because the value of Y is not available for some people. Instead we can compute $\Pr[Y=1|A=a, C=0]$ where C indicates whether the subject was lost (1: yes, 0: no). This restriction to subjects with $C=0$ is problematic because subjects that were lost ($C=1$) may not be exchangeable with subjects who remained through the end of the study ($C=0$). For example, if subjects who did not receive a transplant ($A=0$) and who had a more severe disease decide to leave the study, then the risk $\Pr[Y=1|A=0, C=0]$ among those remaining in the study would be lower than the risk $\Pr[Y=1|A=0]$ among those originally assigned to medical treatment. Our association measure $\Pr[Y=1|A=1, C=0]/\Pr[Y=1|A=0, C=0]$ would not generally equal the effect measure $\Pr[Y_{a=1}=1]/\Pr[Y_{a=0}=1]$.
- *Non-compliance*. Subjects may not adhere to the assigned treatment. Let A be the exposure to which subjects were randomly assigned, and B the exposure they actually received. Suppose some subjects that had been assigned to medical treatment ($A=0$) obtained a heart transplant outside of the study ($B=1$). In an "intention to treat" analysis, we compute $\Pr[Y=1|A=a]$, which equals $\Pr[Y_a=1]$. However, we are not interested in the causal effect of assignment A, a misclassified version of the true exposure B, but on the causal effect of B itself. The alternative "as treated" approach—using $\Pr[Y=1|B=b]$ for causal inference—is problematic. For example, if the most severely ill subjects in the $A=0$ group seek a heart transplant ($B=1$) outside of the study, then the group $B=1$ would include a higher proportion of severely ill subjects than the group $B=0$. The groups $B=1$ and $B=0$ would not be exchangeable—that is, $\Pr[Y=1|B=b] \neq \Pr[Y_b=1]$. In the presence of non-compliance, an intention to treat analysis guarantees exchangeability of the groups defined by a misclassified exposure (the original assignment), whereas an as treated analysis guarantees a correct classification of exposure but not exchangeability of the groups defined by this exposure. However, the intention to treat analysis is often preferred because, unlike the as treated analysis, it provides an unbiased association measure if the sharp causal null hypothesis holds for the exposure B.
- *Unblinding*. When the study subjects are aware of the treatment they receive (as in our heart transplant study), they may change their behaviour accordingly. For example, those who received a transplant may change their diet to keep their new heart healthy. The equality $\Pr[Y=1|A=a] = \Pr[Y_a=1]$ still holds, but now the causal effect of A combines the effects of the transplant and the dietary change. To avoid this problem, knowledge of the level of exposure assigned to each group is withheld from subjects and their doctors (they are "blinded"), when possible. The goal is to ensure that the whole effect, if any, of the exposure assignment A is solely attributable to the exposure received B (the heart transplant in our example). When this goal is achieved, we say that the *exclusion restriction* holds—that is, $Y_{a=0,b} = Y_{a=1,b}$ for all subjects and all values b and, specifically, for the value B observed for each subject. In non-blinded studies, or when blinding does not work (for example, the well known side effects of a treatment make apparent who is taking it), the exclusion restriction cannot be guaranteed, and therefore the intention to treat analysis may not yield an unbiased association measure even under the sharp causal null hypothesis for exposure B.

In summary, the fact that exchangeability $Y_a \coprod A$ holds in a well designed randomised experiment does not guarantee an unbiased estimate of the causal effect because: *i*) Y may not be measured for all subjects (loss to follow up), *ii*) A may be a misclassified version of the true exposure (non-compliance), and *iii*) A may be a combination of the exposure of interest plus other actions (unblinding). Causal inference from randomised studies in the presence of these problems requires similar assumptions and analytical methods as causal inference from observational studies.

Leaving aside these methodological problems, randomised experiments may be unfeasible because of ethical, logistic, or financial reasons. For example, it is questionable that an ethical committee would have approved our heart transplant study. Hearts are in short supply and society favours assigning them to subjects who are more likely to benefit from the transplant, rather than assigning them randomly among potential recipients. Randomised experiments of harmful exposures (for example, cigarette smoking) are generally unacceptable too. Frequently, the only option is conducting observational studies in which exchangeability is not guaranteed.

BIBLIOGRAPHICAL NOTES

Hume[1] hinted a counterfactual theory of causation, but the application of counterfactual theory to the estimation of causal effects via randomised experiments was first formally proposed by Neyman.[2] Rubin[3 4] extended Neyman's theory to the estimation of the effects of fixed exposures in randomised *and* observational studies. Fixed exposures are exposures that either are applied at one point in time only or never change over time. Examples of fixed exposures in epidemiology

are a surgical intervention, a traffic accident, a one dose immunisation, or a medical treatment that is continuously administered during a given period regardless of its efficacy or side effects. Rubin's counterfactual model has been discussed by Holland and others.[5]

Robins[6 7] proposed a more general counterfactual model that permits the estimation of total and direct effects of fixed *and* time varying exposures in longitudinal studies, whether randomised or observational. Examples of time varying exposures in epidemiology are a medical treatment, diet, cigarette smoking, or an occupational exposure. For simplicity of presentation, our article was restricted to the effects of fixed exposures. The use of the symbol $\amalg$ to denote independence was introduced by Dawid.[8]

ACKNOWLEDGEMENTS

The author is deeply indebted to James Robins for his contributions to earlier versions of this manuscript.

Funding: NIH grant KO8-AI-49392

Conflicts of interest: none declared.

APPENDIX

A1 SAMPLING VARIABILITY

Our descriptions of causal effect and exchangeability have relied on the idea that we somehow collected information from all the subjects in the population of interest. This simplification has been useful to focus our attention on the conceptual aspects of causal inference, by keeping them separate from aspects related to random statistical variability. We now extend our definitions to more realistic settings in which random variability exists.

Many real world studies are based on samples of the population of interest. The first consequence of working with samples is that, even if the counterfactual outcomes of all subjects in the study were known, one cannot obtain the exact proportion of subjects in the population who had the outcome under exposure value a—that is, the probability $\Pr[Y_{a=0}=1]$ cannot be directly computed. One can only estimate this probability. Consider the subjects in table 2. We have previously viewed them as forming a 20 person population. Let us now view them as a random sample of a much larger population. In this sample, the proportion of subjects who would have died if unexposed is $\hat{\Pr}[Y_{a=0}=1] = 10/20 = 0.5$, which does not have to be exactly equal to the proportion of subjects who would have died if the entire population had been unexposed, $\Pr[Y_{a=0}=1]$. We use the sample proportion $\hat{\Pr}[Y_a=1]$ to estimate the population probability $\Pr[Y_a=1]$. (The ''hat'' over Pr indicates that $\hat{\Pr}[Y_a=1]$ is an estimator.) We say that $\hat{\Pr}[Y_a=1]$ is a *consistent estimator* of $\Pr[Y_a=1]$ because the larger the number of subjects in the sample, the smaller the difference between $\hat{\Pr}[Y_a=1]$ and $\Pr[Y_a=1]$ is expected to be. In the long run (that is, if the estimator is applied to infinite samples of the population), the mean difference is expected to become zero.

There is a causal effect of A on Y in such population if $\Pr[Y_{a=1}=1] \neq \Pr[Y_{a=0}=1]$. This definition, however, cannot be directly applied because the population probabilities $\Pr[Y_a=1]$ cannot be computed, but only consistently estimated by the sample proportions $\hat{\Pr}[Y_a=1]$. Therefore, one cannot conclude with certainty that there is (or there is not) a causal effect. Rather, standard statistical procedures are needed to test the causal null hypothesis $\Pr[Y_{a=1}=1] = \Pr[Y_{a=0}=1]$ by comparing $\hat{\Pr}[Y_{a-1}=1]$ and $\hat{\Pr}[Y_{a-1}=1]$, and to compute confidence intervals for the effect measures. The availability of data from only a sample of subjects in the population, even if the values of all their counterfactual outcomes were known, is the first reason why statistics is necessary in causal inference.

The previous discussion assumes that one can have access to the values of both counterfactual outcomes for each subject in the sample (as in table 2), whereas in real world studies one can only access the value of one counterfactual outcome for each subject (as in table 3). Therefore, whether one is working with the whole population or with a sample, neither the probability $\Pr[Y_a=1]$ or its consistent estimator $\hat{\Pr}[Y_a=1]$ can be directly computed for any value a. Instead, one can compute the sample proportion of subjects that develop the outcome among the exposed, $\hat{\Pr}[Y=1|A=1] = 7/13$, and among the unexposed, $\hat{\Pr}[Y=1|A=0] = 3/7$. There are two major conceptualisations of this problem:

(1) The population of interest is near infinite and we hypothesise that all subjects in the population are randomly assigned to either $A=1$ or $A=0$. Exchangeability of the exposed and unexposed would hold in the population—that is, $\Pr[Y_a=1] = \Pr[Y=1|A=a]$. Now we can see our sample as a random sample from this population where exposure is randomly assigned. The problem boils down to standard statistical inference with the sample proportion $\hat{\Pr}[Y=1|A=a]$ being a consistent estimator of the population probability $\Pr[Y=1|A=a]$. This is the simplest conceptualisation.

(2) Only the subjects in our sample, not all subjects in the entire population, are randomly assigned to either $A=1$ or $A=0$. Because of the presence of random sampling variability, we do not expect that exchangeability will exactly hold in our sample. For example, suppose that 100 subjects are randomly assigned to either heart transplant ($A=1$) or medical treatment ($A=0$). Each subject can be classified as good or bad prognosis at the time of randomisation. We say that the groups $A=0$ and $A=1$ are exchangeable if they include exactly the same proportion of subjects with bad prognosis. By chance, it is possible that 17 of the 50 subjects assigned to $A=1$ and 13 of the 50 subjects assigned to $A=0$ had bad prognosis. The two groups are not exactly exchangeable. However, if we could draw many additional 100 person samples from the population and repeat the randomised experiment in each of these samples (or, equivalently, if we could increase the size of our original sample), then the imbalances between the groups $A=1$ and $A=0$ would be increasingly attenuated. Under this conceptualisation, the sample proportion $\hat{\Pr}[Y=1|A=a]$ is a consistent estimator of $\hat{\Pr}[Y_a=1]$, and $\hat{\Pr}[Y_a=1]$ is a consistent estimator of the population proportion $\Pr[Y_a=1]$ if our sample is a random sample of the population of interest. This is the most realistic conceptualisation.

Under either conceptualisation, standard statistical procedures are needed to test the causal null hypothesis $\Pr[Y_{a=1}=1] = \Pr[Y_{a=0}=1]$ by comparing $\hat{\Pr}[Y=1|A=1]$ and $\hat{\Pr}[Y=1|A=0]$, and to compute confidence intervals for the estimated association measures, which are consistent estimators of the effect measures. The availability of the value of only one counterfactual outcome for each subject, regardless of whether all subjects in the population of interest are or are not included the study (and regardless of which conceptualisation is used), is the second reason why statistics is necessary in causal inference.

A2 GENERALISATIONS

A2.1 Definition of causal effect

We defined causal effect of the exposure on the outcome, $\Pr[Y_{a=1}=1] \neq \Pr[Y_{a=0}=1]$, as a difference between the counterfactual risk of the outcome had everybody in the

population of interest been exposed and the counterfactual risk of the outcome had everybody in the population been unexposed. In some cases, however, investigators may be more interested in the causal effect of the exposure in a subset of the population of interest (rather than the effect in the entire population). This causal effect is defined as a contrast of counterfactual risks in that subset of the population of interest.

A common choice is the subset of the population comprised by the subjects that were actually exposed. Thus, we can define the *causal effect in the exposed* as $\Pr[Y_{a=1}=1|A=1] \neq \Pr[Y_{a=0}=1|A=1]$ or, by definition of counterfactual outcome, $\Pr[Y=1|A=1] \neq \Pr[Y_{a=0}=1|A=1]$. That is, there is a causal effect in the exposed if the risk of the outcome among the exposed subjects in the population of interest does not equal the counterfactual risk of the outcome had the exposed subjects in the population been unexposed. The causal risk difference in the exposed is $\Pr[Y=1|A=1]-\Pr[Y_{a=0}=1|A=1]$, the causal risk ratio in the exposed is $\Pr[Y=1|A=1]/\Pr[Y_{a=0}=1|A=1]$, and the causal odds ratio in the exposed is $(\Pr[Y=1|A=1]/\Pr[Y=0|A=1])/(\Pr[Y_{a=0}=1|A=1]/\Pr[Y_{a=0}=0|A=1])$.

The causal effect in the entire population can be computed under the condition that the exposed and the unexposed are exchangeable—that is, $Y_a \coprod A$ for $a=0$ and $a=1$. On the other hand, the causal effect in the exposed can be computed under the weaker condition that the exposed and the unexposed are exchangeable had they been unexposed—that is, $Y_a \coprod A$ for $a=0$ only. Under this weaker exchangeability condition, the risk of the outcome under no exposure is equal for the exposed and the unexposed: $\Pr[Y_{a=0}=1|A=1] = \Pr[Y_{a=0}=1|A=0]$. By definition of a counterfactual outcome $\Pr[Y_{a=0}=1|A=0] = \Pr[Y=1|A=0]$. Therefore, when the exposed and unexposed are exchangeable under $a=0$, $\Pr[Y_{a=0}=1|A=1] = \Pr[Y_{a=0}=1|A=0] = \Pr[Y=1|A=0]$. We decided to restrict our discussion to the causal effect in the entire population and not to the causal effect in the exposed because the latter cannot be directly generalised to time varying exposures.

A2.2 Non-dichotomous outcome and exposure

The definition of causal effect can be generalised to non-dichotomous exposure A and outcome Y. Let $E[Y_a]$ be the mean counterfactual outcome had all subjects in the population received exposure level a. For discrete outcomes, the expected value $E[Y_a]$ is defined as the weighted sum $\sum_y y p_{Y_a}(y)$ over all possible values y of the random variable Y_a, where $p_{Y_a}(\cdot)$ is the probability mass function of Y_a—that is, $p_{Y_a}(y) = \Pr[Y_a = y]$. For continuous outcomes, the expected value $E[Y_a]$ is defined as the integral $\int y\, f_{Y_a}(y)dy$ over all possible values y of the random variable Y_a, where $f_{Y_a}(\cdot)$ is the probability density function of Y_a. A common representation of the expected value for discrete and continuous outcomes is $E[Y_a] = \int y\, dF_{Y_a}(y)$, where $F_{Y_a}(\cdot)$ is the cumulative density function (cdf) of the random variable Y_a.

We say that there is a population *average causal effect* if $E[Y_a] \neq E[Y_{a'}]$ for any two values a and a'. In ideal randomised experiments, the expected value $E[Y_a]$ can be consistently estimated by the average of Y among subjects with $A=a$. For dichotomous outcomes, $E[Y_a] = \Pr[Y_a=1]$.

The average causal effect is defined by the contrast of $E[Y_a]$ and $E[Y_{a'}]$. When we talk of "the causal effect of heart transplant (A)" we mean the contrast between "receiving a heart transplant ($a=1$)" and "not receiving a heart transplant ($a=0$)." In this case, we may not need to be explicit about the particular contrast because there are only two possible actions, and therefore only one possible contrast. But for non-dichotomous exposure variables A, the particular contrast of interest needs to be specified. For example, "the causal effect of aspirin" is meaningless unless we specify that the contrast of interest is, say, "taking 150 mg of aspirin daily for five years" compared with "not taking aspirin". Note that this causal effect is well defined even if counterfactual outcomes under interventions other than those involved in the causal contrast of interest are not well defined or even do not exist (for example, "taking 1 kg of aspirin daily for five years").

The average causal effect, defined as a contrast of means of counterfactual outcomes, is the most commonly used causal effect. However, the causal effect may also be defined by a contrast of, say, medians, variances, or cdfs of counterfactual outcomes. In general, the causal effect can be defined as a contrast of any functional of the distributions of counterfactual outcomes under different exposure values. The causal null hypothesis refers to the particular contrast of functionals (means, medians, variances, cdfs, ...) used to define the causal effect.

A2.3 Non-deterministic counterfactual outcomes

We have defined the counterfactual outcome Y_a as the subject's outcome had he experienced exposure value a. For example, in our first vignette, Zeus would have died if treated and would have survived if untreated. This definition of counterfactual outcome is deterministic because each subject has a fixed value for each counterfactual outcome, for example, $Y_{a=1}=1$ and $Y_{a=0}=0$ for Zeus. However, we could imagine a world in which Zeus has certain probability of dying, say $Q_{Y_{a=1}}(1)=0.9$, if treated and certain probability of dying, say $Q_{Y_{a=0}}(1)=0.1$, if untreated. This is a non-deterministic or stochastic definition of counterfactual outcome because the probabilities $Q_{Y_a}(\cdot)$ are not zero or one. In general, the probabilities $Q_{Y_a}(\cdot)$ vary across subjects (that is, they are random) because not all subjects are equally susceptible to develop the outcome. For discrete outcomes, the expected value $E[Y_a]$ is then defined as the weighted sum $\sum_y y\, p_{Y_a}(y)$ over all possible values y of the random variable Y_a, where the probability mass function $p_{Y_a}(\cdot) = E[Q_{Y_a}(\cdot)]$.

More generally, a non-deterministic definition of counterfactual outcome does not attach some particular value of the random variable Y_a to each subject, but rather a statistical distribution $\Theta_{Y_a}(\cdot)$ of Y_a. The deterministic definition of counterfactual outcome implies that the cdf $\Theta_{Y_a}(y)$ can only take values 0 or 1 for all y. The use of random distributions of Y_a (that is, distributions that may vary across subjects) to allow for non-deterministic counterfactual outcomes does not imply any modification in the definition of average causal effect or the methods used to estimate it. To show this, first note that $E[Y_a] = E[E[Y_a|\Theta_{Y_a}(\cdot)]]$. Therefore, $E[Y_a] = E\left[\int y\, d\Theta_{Y_a}(y)\right] = \int y\, dE[\Theta_{Y_a}(y)] = \int y\, dF_{Y_a}(y)$ because $F_{Y_a}(\cdot) = E[\Theta_{Y_{a,i}}(\cdot)]$. The non-deterministic definition of causal effect is a generalisation of the deterministic definition in which $\Theta_{Y_a}(\cdot)$ is a general cdf that may take values between 0 and 1.

The choice of deterministic compared with non-deterministic counterfactual outcomes has no consequences for the definition of the average causal effect and the point estimation of effect measures based on averages of counterfactual outcomes. However, this choice has implications for the computation of confidence intervals for the effect measures.[9]

A3 NO INTERACTION BETWEEN SUBJECTS

An implicit assumption in our definition of individual causal effect is that a subject's counterfactual outcome under exposure value a does not depend on other subjects' exposure value. This assumption was labelled "no interaction between

units'' by Cox,[10] and ''stable-unit-treatment-value assumption (SUTVA)'' by Rubin.[11] If this assumption does not hold (for example, in studies dealing with contagious diseases or educational programmes), then individual causal effects cannot be identified by using the hypothetical data in table 2. Most methods for causal inference assume that SUTVA holds.

A4 POSSIBLE WORLDS

Some philosophers of science define causal effects using the concept of ''possible worlds.'' The actual world is the way things actually are. A possible world is a way things might be. Imagine a possible world a where everybody receives exposure value a, and a possible world a' where everybody received exposure value a'. The mean of the outcome is $E[Y_a]$ in the first possible world and $E[Y_{a'}]$ in the second one. There is a causal effect if $E[Y_a] \neq E[Y_{a'}]$ and the worlds a and a' are the two worlds closest to the actual world where all subjects receive exposure value a and a', respectively.

We introduced the counterfactual Y_a as the outcome of a certain subject under a well specified intervention that exposed her to a. Some philosophers prefer to think of the counterfactual Y_a as the outcome of the subject in the possible world that is closest to our world and where she was exposed to a. Both definitions are equivalent when the only difference between the closest possible world involved and the actual world is that the intervention of interest took place. The possible worlds' formulation of counterfactuals replaces the difficult problem of specifying the intervention of interest by the equally difficult problem of describing the closest possible world that is minimally different from the actual world. The two main counterfactual theories based on possible worlds, which differ only in details, have been proposed by Stalnaker[12] and Lewis.[13]

REFERENCES

1 **Hume D**. *An enquiry concerning human understanding*. [Reprinted and edited 1993]. Indianapolis/Cambridge: Hacket, 1748.
2 **Neyman J**. On the application of probability theory to agricultural experiments: essay on principles, section 9. *Translated in Statistical Science* 1923, 1990;**5**:465–80.
3 **Rubin DB**. Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology* 1974;**56**:688–701.
4 **Rubin DB**. Bayesian inference for causal effects: the role of randomization. *Annals of Statistics* 1978;**6**:34–58.
5 **Holland PW**. Statistics and causal inference (with discussion). *Journal of the American Statistical Association* 1986;**81**:945–61.
6 **Robins JM**. A new approach to causal Inference in mortality studies with sustained exposure periods_application to control of the healthy worker survivor effect. *Mathematical Modelling* 1986;**7**:1393–512 (errata appeared in *Computers and Mathematics with Applications* 1987;**14**:917–21).
7 **Robins JM**. Addendum to ''A new approach to causal inference in mortality studies with sustained exposure periods_application to control of the healthy worker survivor effect''. *Computers and Mathematics with Applications* 1987;**14**:923–45. (errata appeared in *Computers and Mathematics with Applications* 1987;**18**:477).
8 **Dawid AP**. Conditional independence in statistical theory (with discussion). *Journal of the Royal Statistical Society B* 1979;**41**:1–31.
9 **Robins JM**. Confidence intervals for causal parameters. *Stat Med* 1988;**7**:773–85.
10 **Cox DR**. *Planning of experiments*. New York: Wiley, 1958.
11 **Rubin DB**. Discussion of ''Randomized analysis of experimental data: the Fisher randomization test'' by Basu D. *Journal of the American Statistical Association* 1980;**75**:591–3.
12 **Stalnaker RC**. A theory of conditionals. In: Rescher N, ed. *Studies in logical theory*. Oxford: Blackwell, 1968. [Reprinted in Jackson F, ed. *Conditionals*. Oxford: Oxford University Press, 1991.]
13 **Lewis D**. *Counterfactuals*. Oxford: Blackwell, 1973.