# Causal inference from experiment and observation

Marcel Zwahlen, Geogia Salanti

Institute of Social and Preventive Medicine, University of Bern, Bern, Switzerland
**Correspondence to** Dr Marcel Zwahlen, Institute of Social and Preventive Medicine, University of Bern, Finkenhubelweg 11, Bern CH 3012, Switzerland; marcel.zwahlen@ispm.unibe.ch

## ABSTRACT

Results from well-conducted randomised controlled studies should ideally inform on the comparative merits of treatment choices for a health condition. In the absence of this, one attempts to use evidence from the impact of treatment when administered according to decisions of the physicians and the patients (observational evidence). Naïve comparisons between treatment options using observational evidence will lead to biased results. Under certain conditions, however, it is possible to obtain valid estimates of the comparative merits of different treatments from observational data. Causal inference can be conceptualised as a framework aiming to provide valid information about causal effects of treatments using observational evidence. It can be viewed as a missing data problem in which each patient has two outcomes: the observed outcome under the treatment actually received and a counterfactual (unobserved) outcome had the patient received a different treatment. Methodological developments over the last decades clarified the appropriate conditions and methods to obtain valid comparisons. This article provides an introduction to some of these methods.

## INTRODUCTION

### Making causal statements about relative treatment effects

Let us start with a simple example by presenting the following (fictional) information on 800 patients with acute depression who received treatment A and another 800 patients with acute depression who got treatment B (table 1) for a complete 6-month follow-up.

We are now invited to make a statement which treatment is more efficacious, that is, is associated with less chances of relapse. As we had previously attended some courses in clinical epidemiology, we refuse to answer directly the question. Instead we ask what type of study this was that produced these data.

We want to know this because our answer depends on the study design.

- If the data are coming from a randomised trial with complete adherence and complete 6-month follow-up, we will probably say that treatment A is more efficacious.
- If the data result from two different clinics in which patients with acute depression were treated, then we would not be sure what to prefer as the difference between the two clinics might be confounded by the differences in patient characteristics across the two clinics. We therefore would ask for more information, for example, the distributions of the initial depression severity or concomitant conditions in these two groups of patients.

Ideally, we would like to know the answers to the following two questions:

- For those who had received treatment A, how many would have relapsed if they had received treatment B?
- For those who had received treatment B, how many would have relapsed if they had received treatment A?

We would then clearly prefer treatment A, if for all 1600 patients, the 6-month relapse rate under treatment A is lower than under treatment B. This statement is an example of a causal statement. We could be even more precise in our causal statement: the 6-month relapse rate under treatment A would be 5.1% lower than under treatment B, if we can argue that the relapse rate observed for those who received treatment A reflects the relapse rate if everyone would have received treatment A and similarly for those who received treatment B (and ignoring issues about statistical uncertainty). But of course, we are now discussing relapse results of hypothetical situations that we will never observe. The idea to discuss hypothetical results from hypothetical situations was introduced several decades ago by Fischer, Neyman and Rubin[1–3] and more recently formalised by Pearl, Hernan and Robins.[4 5] Confronted with all these rather long 'if' statements , we start to feel the need for some more formal notation[5 6]:

- Let Y denote the binary outcome for a patient having relapsed within 6 months (yes; Y=1, no; Y=0).
- Let Tr denote the treatment someone has actually received or *could have* received (Tr=A or Tr=B).
- Let $Y_{Tr=A}$ denote the (*potential or observed*) outcome for a patient if (s)he would have received treatment A, similarly $Y_{Tr=B}$ denotes the (*potential or observed*) outcome for a patient if (s)he would have received treatment B. In some publications subscripts are used,[7] as here, sometimes superscripts ($Y^{Tr=A}$ or $Y^{Tr=B}$) denote the potential outcomes of an individual.[5]
- Let Pr[] denote the probability that something happened or the proportion of situations in which something happened, so Pr[Y=1 | Tr=A] denotes the proportion of patients who relapsed among those who actually received treatment A.

The outcomes not observed, that is, $Y_{Tr=B}$ in patients who received A and $Y_{Tr=A}$ in patients who received B, are called counterfactual outcomes. Table 2 illustrates this; for those patients who received treatment A, we are able to observe $Y_{Tr=A}$, but not $Y_{Tr=B}$; similarly for those who received treatment B, we observe $Y_{Tr=B}$ but not $Y_{Tr=A}$. The information we have allows us to calculate the proportion of relapses among patients indeed receiving treatment A and those indeed receiving treatment B, that is, Pr[Y=1 | Tr=A] and Pr[Y=1 | Tr=B].

We can conceptually define an individual causal effect for each of the 1600 persons in table 1. For example, we can define the difference between the outcomes under different treatments $Y_{Tr=B} - Y_{Tr=A}$, or assume that there is no difference in the treatment outcome, that is, $Y_{Tr=B} = Y_{Tr=A}$ for each of the 1600 patients. Again, note that we are not able to observe any of these individual causal effects. However, perhaps, we can make statements about the population causal effect. By that, we

**Table 1** Observed 6-month relapse for patients with acute depression receiving either treatment A or B

| | Treatment A | | | Treatment B | | |
|---|---|---|---|---|---|---|
| | Number of patients | Relapses | Per cent with relapse | Number of patients | Relapses | Per cent with relapse |
| Total | 800 | 40 | 5.0 | 800 | 81 | 10.1 |

**Table 2** Data and counterfactual outcomes for the 10 first patients with treatments (Tr) A and B

| Person | Treatment received (Tr) | Outcome Y observed | $Y_{Tr=A}$ | $Y_{Tr=B}$ |
|---|---|---|---|---|
| 1 | A | 0 | 0 | ? |
| 2 | B | 1 | ? | 1 |
| 3 | B | 0 | ? | 0 |
| 4 | A | 1 | 1 | ? |
| 5 | A | 0 | 0 | ? |
| 6 | A | 0 | 0 | ? |
| 7 | B | 1 | ? | 1 |
| 8 | B | 0 | ? | 0 |
| 9 | B | 0 | ? | 0 |
| 10 | B | 0 | ? | 0 |

The dichotomous outcome (relapse) is denoted with Y. The counterfactual outcomes are denoted with '?'.

mean the proportion of patients who would relapse if *all* 1600 would have received treatment B, compared with the relapse risk if *all* patients would have received treatment A, or (using our notation) estimate $Pr[Y_{Tr=B}=1] - Pr[Y_{Tr=A}=1]$. Unlike individual causal effects, it is possible—under certain conditions—to estimate population causal effects.

## EXCHANGEABILITY ALLOWS TO ESTIMATE POPULATION CAUSAL EFFECTS

Let us revisit the situation in table 1 and assume the table gives the results of a well conducted randomised study with complete follow-up. Apparently, the randomisation was 1:1 as 800 patients received treatment A and 800 patients received treatment B. The average of the counterfactual outcomes of the 800 patients who received A had they received B is simply the average observed outcome of those 800 patients who did receive B; in other words, the two groups, A and B, are exchangeable. Therefore, those who received treatment A are a perfect random sample of all 1600 patients, and the relapse rate we observe among those who received treatment A, $Pr[Y=1 \mid Tr=A]$, estimates (up to sampling uncertainty) what would have happened if all would have received treatment A, that is, $Pr[Y_{Tr=A}=1]$. The same argument can be made for those who indeed received treatment B. They are a perfect random sample of all and therefore $Pr[Y=1 \mid Tr=B]=Pr[Y_{Tr=B}=1]$. The random treatment assignment allows to estimate (on average) what would have been if those receiving A would have received B by looking at those who actually received B. We get an *average* estimate for the question marks in table 2 by using the treatment results from the other group.

The term "exchangeability" refers to the fact that the relapse risk under the possible treatment choices A or B among those who actually received A (ie, $Pr[Y_{Tr=A}=1 \mid Tr=A]$ and $Pr[Y_{Tr=B}=1 \mid Tr=A]$) equals the risk under the possible treatment choices A or B among those who actually received B (ie, $Pr[Y_{Tr=A}=1 \mid Tr=B]$ and $Pr[Y_{Tr=B}=1 \mid Tr=B]$). Randomisation produces exchangeability, and hence functions of the observed average outcomes can be interpreted as causal effects of the treatments.

## LACK OF EXCHANGEABILITY

Exchangeability would be clearly violated if participants differ considerably across the two treatments in characteristics that are related to the outcome of interest. Such a case can occur when one treatment is preferentially been given more often to patients with more severe depression. Let us assume both treatments are equally effective, that is, the population causal risk difference is zero, that is, $Pr[Y_{Tr=A}=1]=Pr[Y_{Tr=B}=1]$. Treatment B is given exclusively in clinic B that receives patients within more severe symptoms who have a higher risk to relapse after treatment. This then leads to a higher observed relapse rate in those who receive treatment B than in those who receive treatment A (administrated in clinic A), that is, $Pr[Y=1 \mid Tr=B] > Pr[Y=1 \mid Tr=A]$, and the observed difference does not correspond to the population causal risk difference.

Let us suppose that the data in table 1 reflect such an observational study where patients are treated in two different clinics and admission to each of the clinics depends on patient characteristics. Table 3 now presents the 6-month relapse results for the two groups of patients stratified by sex, age group and severity of depression symptoms at study enrolment. Those treated in clinic B are more often males (510 of 800) compared with clinic A (400 of 800), and in clinic B, patients are on average older (510 were 60+ years) compared with patients in clinic A (280 were 60+ years), and finally clinic B had more patients with severe depression symptoms (400 patients) than clinic A (200 patients). Within each of the eight subgroups (by sex, age and symptoms severity level), we observe the same relapse rate between the two clinics.

We see that the information looks different when we stratify the data into these subgroups. Can we conclude that relapse rate after treatment with A is equivalent to that with B? Being equivalent would mean that if all had been treated with A, we would expect the same results as if all would have been treated with B. To be able to conclude this equivalence, we need to assume that within these eight subgroups exchangeability is fulfilled. This is equivalent to assume that within the eight subgroups, the 'assignment' to clinic A or B is 'as randomised' (although not in 1:1 randomisation ratio but in a ratio that is changing from subgroup to subgroup). This also implies that we are assuming that there is no further important variable that was not assessed. This also known as the assumption of no unmeasured confounding.

**Table 3** Observed 6-month relapse for patients with depression receiving either treatment A or B stratified by sex, age group and severity of symptoms at study enrolment

| | | | Treatment with A in clinic A | | | Treatment with B in clinic B | | |
|---|---|---|---|---|---|---|---|---|
| | | | Number of patients | Relapses | Per cent with relapse | Number of patients | Relapses | Per cent with relapse |
| Total | Age | Severity | 800 | 40 | 5.0 | 800 | 81 | 10.1 |
| Men | <60 | Low | 200 | 4 | 2.0 | 50 | 1 | 2.0 |
| | <60 | High | 60 | 6 | 10.0 | 100 | 10 | 10.0 |
| | 60+ | Low | 100 | 5 | 5.0 | 200 | 10 | 5.0 |
| | 60+ | High | 40 | 10 | 25.0 | 160 | 40 | 25.0 |
| Women | <60 | Low | 200 | 2 | 1.0 | 100 | 1 | 1.0 |
| | <60 | High | 60 | 3 | 5.0 | 40 | 2 | 5.0 |
| | 60+ | Low | 100 | 4 | 4.0 | 50 | 2 | 4.0 |
| | 60+ | High | 40 | 6 | 15.0 | 100 | 15 | 15.0 |

## HOW TO OBTAIN THE POPULATION CAUSAL EFFECT FROM OBSERVATIONAL DATA

First, we need to assume exchangeability within the subgroups that is called conditional exchangeability.[7] Once we assume conditional exchangeability, we have different choices for comparing the relapse rates between treatments. One option is to use a logistic regression model for relapse including as predictors age, sex and severity in addition to treatment. The coefficient for treatment of such a logistic regression model makes implicit comparisons of patients of the same sex, age and depression severity. Some would suggest to use Mantel-Haenszel methods[8] or a propensity score matching procedure.[9] Another approach is what basic epidemiology books describe as direct standardisation[7]: calculate the expected relapse in all 1600 patients (in the eight subgroups) first using the observed relapse of clinic A and a then using the observed relapse of clinic B. Then compare the two expected relapse rates in all 1600 patients. The difference between the two expected relapse rates under A and B is the population causal effect (under the assumption of conditional exchangeability within the subgroups). A bit more work is then needed to obtain an appropriate 95% CI for the causal effect. Instead of doing the direct standardisation calculation steps as just described, we could approach the calculations by using the so-called inverse probability of treatment weights.

## USE OF INVERSE PROBABILITY OF TREATMENT WEIGHTS

The idea behind the use of inverse probability of treatment weights (IPTW) is to create two 'pseudopopulations' of patients of the same total size as the one observed. In one pseudopopulation, all receive treatment A; in the other, all receive treatment B. Then, the probability (risk) of relapse in the two pseudopopulations will be compared.

Let us illustrate the idea by looking only at one subgroup: that of men of aged 60+ years with high severity depression in table 3. In total, there are 200 participants in this subgroup: 40 treated with A and 160 treated with B. So, a man aged 60+ years with high severity depression has 1 in 5 probability to be treated with A and 4 in 5 to be treated with B. To calculate what would have happened to 200 patients if they were all treated with A, we multiply the 40 patients (and their 10 observed relapses) indeed treated at clinic A by a factor of 5, which is the inverse of the probability to receive A. So, out of 200 patients, 50 would relapse if they were all treated with A. To calculate the expected number of relapses had all 200 patients received B, we multiply the observed relapse rate in B (40 in 160) with a factor 5/4=1.25 (the inverse of the probability to receive B). So, again 50 patients would relapse if all 200 had been treated with B.

Therefore, we can get the results of the direct standardisation by the following steps:

1. For each of the 8 subgroups we assume exchangeability and calculate the probability to receive the treatment (s)he has indeed received. Denote these with Pr[Tr as received | subgroup].
2. Calculate IPTW=1/Pr[Tr as received | subgroup].
3. Calculate the observed event rate within each subgroup Pr[Y=1 | Tr as received in subgroup].
4. Within each subgroup, multiply IPTW with the number of patients in the subgroup and with the number of events. This will reconstruct what we expect if all would have had treatment A compared with if all would have had treatment B.
5. Sum the events and number of patients across subgroups in the pseudopopulations for A and B. Use these numbers to estimate the causal relative treatment effect.
6. Obtain a 95% CI for the causal relative treatment effect by using robust standard errors.[5 10 11]

Table 4 gives the IPTW weights for each of the eight subgroups that are used to create the two pseudopopulations. One hundred and sixty-one patients would relapse out of 1600 patients who could have received A; the same for B and hence the risk difference is zero.

The steps outlined above are easily done using standard statistical software. The probabilities in step 1 can be obtained from a logistic regression with receiving treatment A (or B) as the outcome (see online supplementary material 1 on how this can be done in Stata). Step

**Table 4** Data, inverse probability of treatment weights (IPTW) and pseudopopulations for treatments A and B

| | | | Observed data | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | | Treatment with A | | | Treatment with B | | | |
| Sex | Age (years) | Severity | Patients | Relapses | IPTW weights | Patients | Relapses | IPTW weights | Total |
| Men | <60 | Low | 200 | 4 | 1.25 | 50 | 1 | 5 | 250 |
| | <60 | High | 60 | 6 | 2.67 | 100 | 10 | 1.6 | 160 |
| | 60+ | Low | 100 | 5 | 3 | 200 | 10 | 1.5 | 300 |
| | 60+ | High | 40 | 10 | 5 | 160 | 40 | 1.25 | 200 |
| Women | <60 | Low | 200 | 2 | 1.5 | 100 | 1 | 3 | 300 |
| | <60 | High | 60 | 3 | 1.67 | 40 | 2 | 2.5 | 100 |
| | 60+ | Low | 100 | 4 | 1.5 | 50 | 2 | 3 | 150 |
| | 60+ | High | 40 | 6 | 3.5 | 100 | 15 | 1.4 | 140 |
| | | All | 800 | 40 | | 800 | 81 | | 1600 |
| | | | Pseudopopulations | | | | | | |
| | | | If all patients were treated with A | | | If all patients were treated with B | | | |
| Men | <60 | Low | 250 | 5 | | 250 | 5 | | 500 |
| | <60 | High | 160 | 16 | | 160 | 16 | | 320 |
| | 60+ | Low | 300 | 15 | | 300 | 15 | | 600 |
| | 60+ | High | 200 | 50 | | 200 | 50 | | 400 |
| Women | <60 | Low | 300 | 3 | | 300 | 3 | | 600 |
| | <60 | High | 100 | 5 | | 100 | 5 | | 200 |
| | 60+ | Low | 150 | 6 | | 150 | 6 | | 300 |
| | 60+ | High | 140 | 21 | | 140 | 21 | | 280 |
| | | All | 1600 | 121 | | 1600 | 121 | | 3200 |

**Table 5** Comparing relapse rate under treatment B with that under treatment A using different analytical approaches to estimate causal ORs while accounting for confounding by sex, age and severity of symptoms

| Analytical approach | OR for relapse (B vs A) and 95% CI |
|---|---|
| Logistic regression for relapse including only hospital (no adjustment) | 2.14 (1.45 to 3.17) |
| Logistic regression for relapse including sex, age and severity independently | 0.98 (0.64 to 1.52) |
| Logistic regression for relapse including sex, age, severity with all 2-way interactions between sex, age and severity | 1.0 (0.65 to 1.55) |
| IPTW weighted analysis with weights constructed with sex, age and severity independently in the model for the defining the weights | 0.99 (0.65 to 1.51) |
| IPTW weighted analysis with weights constructed with all two-way interactions between sex, age and severity in the model for the defining the weights | 1.0 (0.65 to 1.53) |

IPTW, inverse probability of treatment weight.

2 is simple, and steps 3 and 6 can be done again with a logistic regression or other generalised linear models that allow for using robust standard errors.[5 10 11]

Table 5 presents the results from the different approaches to analyse the data of table 3 and obtain an OR for relapse between the two treatments. A naïve crude analysis, not accounting for the different patient profiles in clinic A and B, results in a clearly higher odds for relapse in clinic B compared with clinic A. The various ways to account for the differences in patient characteristics and obtain causal effects do not show remarkable differences in the estimated OR.

## ADVANTAGES IN USING INVERSE PROBABILITY WEIGHTING IN ESTIMATING CAUSAL EFFECTS

If IPTW results are comparable with those obtained from a logistic model, why is the use of IPTW weights (or direct standardisation) needed in practice? The reason is that the use of IPTW weights can be extended to situations in which standard regression models will not allow to reconstruct an 'as randomised' situation, especially if time-dependent confounding exists.[5 12 13] This happens in observational studies with treatments that vary over time, when the treatment depends on the patients' outcome and when time-dependent confounders are present that are also affected by previous treatments. For example, estimation of the causal effect of timing of starting antiretroviral treatment would be problematic with standard regression approaches; this is because treatment decision in HIV-infected persons are based on the concentration of CD4+ lymphocytes measured in the blood.[14–17] The concentration of CD4+ lymphocytes declines over time in the absence of antiretroviral treatment, and there is a higher mortality risk for lower concentrations. However, the concentration of CD4+ lymphocytes is also affected by previous treatment as effective treatment increases the concentration. Methods based on using IPTW allowed researchers to estimate the causal effect where standard methods may fail to adjust appropriately for the time-changing CD4 concentrations.[12 14]

The use of inverse probability weights can also be extended to address more complex situations, for example, observations with differential follow-up as encountered in a study on patients with stroke in Switzerland.[18] The study attempted to assess the 12-month vital status of all patients with stroke being discharged from clinics in a defined region in Switzerland in 2008.[18] Table 6 shows a simplified version of the data with just one strong prognostic factor for mortality. Patients could be separated into a high and low relapse risk group based on the National Institutes of Health (NIH) stroke scale. Here we have 1000 patients leaving the clinic, and 660 (66%) could be traced for follow-up information. However, availability of follow-up information was not the same in the low-risk and high-risk groups. Follow-up information was available in 90% for the low-risk group patients and in 50% of the high-risk group patients. Mortality risk among those with available follow-up information was 34%, with 20% in the low-risk patients and 50% in the high-risk patients. Clearly, it would be inappropriate to think that this 34% mortality reflects the true mortality rate among all 1000 patients.

So how could one obtain a more realistic mortality estimate? Again, as with the data in table 3, an additional assumption about (conditional) exchangeability is needed. If one is willing to assume that within each risk group the patients with follow-up information are representative of all the patients of that risk group, one would do the following calculation to reconstruct the mortality among all 1000 patients. We expect to have a 20% mortality among all 400 low-risk patients (=80 expected deaths) and a 50% mortality in high-risk patients (=300 expected deaths). In total, we expect 380 deaths among all 1000 patients, that is, a mortality risk of 38%.

We get mathematically exactly the same result (38%) if one would conduct a weighted analysis restricted to the 660 observed patients with available follow-up information, but using risk group specific weights, which are 1.11 (1/0.8) and 2 (1/0.5), derived as the inverse of the probability of available follow-up information. This is what we called an analysis using inverse probability of censoring weights. The advantages of such a weighted approach are twofold. First, it can easily be extended to more than one prognostic variable risk using multivariable logistic regression to construct the weights. Second, as discussed above with the IPTW, in almost all statistical software, it is possible to conduct a weighted analysis and to obtain estimates and robust 95% CI that account for the weighting.[10 11] However, we need to remember and acknowledge the assumption that all relevant prognostic variables have been included in a correct statistical model for calculating the weights. The risk estimate derived from inverse probability of censoring weights might still be biased, if this assumption does not hold. In the Bern stroke patient study, the naïve estimate of the 12-month mortality was 20.6% (95% CI 17.6% to 24.0%). When using inverse probability of censoring weights derived from a logistic regression including eight baseline characteristics (sex, age, NIH stroke scale, diabetes, smoking, hyperlipidaemia, hypertension

**Table 6** The use of inverse probability of censoring weights when assessing 12-month vital status of patients with stroke being discharged from clinics

| Severity of disease or risk category for death | Number of patients leaving the clinic | Patients with available follow-up information | Deaths recorded at follow-up | Per cent dead at follow-up | Probability of having a follow-up (%) | Inverse of the probability of having a follow-up |
|---|---|---|---|---|---|---|
| Low | 400 | 360 | 72 | 20 | 90.0 | 1.11 |
| High | 600 | 300 | 150 | 50 | 50.0 | 2.00 |
| Total | 1000 | 660 | 222 | 34 | | |

and Charlson index), 12-month mortality was estimated at 27.6% (95% CI 23.7% to 31.5%). Finally, inverse probability of treatment weights and inverse probability of censoring weights can jointly be combined to obtain estimates of all patients treated one way or the other with complete follow-up (if assumptions about conditional exchangeability hold up).

## CONCLUDING REMARKS

In this tutorial, we covered a formal definition of population causal effects using arguments on counterfactual outcomes ($Pr[Y_{Tr=A}=1]$ vs $Pr[Y_{Tr=B}=1]$ in case of binary outcome) and explained how the use of inverse probability of treatment weights and inverse probability of censoring weights plus certain exchangeability assumptions allow to calculate the difference or the (odds) ratio of $Pr[Y_{Tr=A}=1]$ and $Pr[Y_{Tr=B}=1]$. So far we discussed how to estimate causal effects when treatments are given at just on time point (like in table 3). Additional methods have been developed over the last two decades to also obtain causal effect estimates when (like in table 3) different long term treatments with incomplete adherence are to be compared.[5 19–22] Furthermore, some refinements to the calculation of weights (like stabilisation of the weights) are often recommended to avoid overly wide 95% CIs (see chapter 12 in ref 5).

Although primarily used in the analysis of observational studies, methods for causal inference are also relevant to the analysis of randomised trials. The fact that exchangeability holds in a well conducted randomised experiment provides no guarantee that the intention-to-treat analysis provides an unbiased estimate of the causal effect.[6 23 24] The outcome may not be measured for all subjects (differential loss to follow-up), the treatment assignment may not reflect actual treatment received (non-compliance) and unblinding of the treatment might result in differential cotreatments plus other actions. Causal inference from randomised studies in the presence of these problems requires similar assumptions and analytical methods as causal inference from observational studies.[23–26]

All the concepts and methods outlined here make the implicit assumption that a subject's counterfactual outcome under one treatment version or exposure value does not depend on other subjects' treatment version. If this assumption does not hold (eg, in studies dealing with infectious diseases or educational reforms), then individual causal effects cannot be properly defined for a given person via the concept of individual potential outcomes ($Y_{Tr=B}$ vs $Y_{Tr=A}$). Some see it as a limitation that the counterfactual approach is conceptualised with 'treatments' or well-defined actions but seems less helpful to other types of scientific questions on causality.[27] Hernan recently responded to this[28]: 'The goal of the potential outcomes framework is not to identify causes or to "prove causality", as it sometimes said. That causality cannot be proven was already forcibly argued by Hume in the 18th century.[29] Rather, quantitative counterfactual inference helps us predict what would happen under different interventions, which requires our commitment to define the interventions of interest'. So the potential outcomes framework helps to organise our discussion and thinking when we—the medical professions, the society—are discussing what is the best way of action.

**Acknowledgements** GS is a Marie Skłodowska-Curie Fellow (Grant Nr MSCA-IF-703254).

**Competing interests** None declared.

**Provenance and peer review** Not commissioned; externally peer reviewed.

► Additional material is published online only. To view please visit the journal online (http://dx.doi.org/10.1136/eb-2017-102859).

doi:10.1136/eb-2017-102859

Received 12 November 2017; Revised 5 December 2017; Accepted 7 December 2017

## REFERENCES

1. **Neyman J,** Dabrowska DM, Speed TP. On the Application of Probability Theory to Agricultural Experiments. *Essay on Principles. Section 9. Statistical Science* 1990;**5**:465–72.
2. **Fischer RA.** *The Design of Experiments*. Macmillian, 1935. (reprinted 1971).
3. **Rubin DB.** Estimating causal effects of treatments in randomized and nonrandomized studies. *J Educ Psychol* 1974;**66**:688–701.
4. **Pearl J.** *Causality*. 2nd edn. New York, USA: Cambridge University Press, 2009.
5. **Hernan MA,** Robins J. Causal Inference. 2017. (forthcoming) https://www.hsph.harvard.edu/miguel-hernan/causal-inference-book/: Chapman & Hall/CRC.
6. **Hernan MA.** A definition of causal effect for epidemiological research. *J Epidemiol Community Health* 2004;**58**:265–71.
7. **Hernan MA,** Robins JM. Estimating causal effects from epidemiological data. *J Epidemiol Community Health* 2006;**60**:578–86.
8. **Mantel N,** Haenszel W, Hammond CE, *et al*. Statistical aspects of the analysis of data from retrospective studies of disease. *J Natl Cancer Inst* 1959;**22**:719–48.
9. **Rosenbaum PR,** Rubin DB. The central role of the propensity score in observational studies for causal effects. *Biometrika* 1983;**70**:41–55.
10. **White H.** A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity. *Econometrica* 1980;**48**:817–38.
11. **White H.** Maximum likelihood estimation of misspecified models. *Econometrica* 1982;**50**:1–26.
12. **Robins JM,** Hernán MA, Brumback B. Marginal structural models and causal inference in epidemiology. *Epidemiology* 2000;**11**:550–60.
13. **Fewell Z,** Hernan MA, Wolfe F, *et al*. Controlling for time-dependent confounding using marginal structural models. *Stata Journal* 2004;**4**:402–20.
14. **Hernán MA,** Brumback B, Robins JM. Marginal structural models to estimate the causal effect of zidovudine on the survival of hiv-positive men. *Epidemiology* 2000;**11**:561–70.
15. **Sterne JAC,** Hernán MA, Ledergerber B, *et al*. Long-term effectiveness of potent antiretroviral therapy in preventing AIDS and death: a prospective cohort study. *The Lancet* 2005;**366**:378–84.
16. **Sterne JA,** May M, Costagliola D, *et al*. Timing of initiation of antiretroviral therapy in AIDS-free HIV-1-infected patients: a collaborative analysis of 18 HIV cohort studies. *Lancet* 2009;**373**:1352–63.
17. **Cain LE,** Logan R, Robins JM, *et al*. When to initiate combined antiretroviral therapy to reduce mortality and AIDS-defining illness in HIV-infected persons in developed countries: an observational study. *Ann Intern Med* 2011;**154**:509–15.
18. **Fischer U,** Mono M-L, Zwahlen M, *et al*. Impact of thrombolysis on stroke outcome at 12 months in a population. *Stroke* 2012;**43**:1039–45.
19. **Hernán MA,** Cole SR, Margolick J, *et al*. Structural accelerated failure time models for survival analysis in studies with time-varying treatments. *Pharmacoepidemiol Drug Saf* 2005;**14**:477–91.
20. **Danaei G,** Rodríguez LAG, Cantero OF, *et al*. Observational data for comparative effectiveness research: an emulation of randomised trials of statins and primary prevention of coronary heart disease. *Stat Methods Med Res* 2013;**22**:70–96.
21. **Hernán MA,** Robins JM. Using big data to emulate a target trial when a randomized trial is not available. *Am J Epidemiol* 2016;**183**:758–64.
22. **Hernán MA,** Sauer BC, Hernández-Díaz S, *et al*. Specifying a target trial prevents immortal time bias and other self-inflicted injuries in observational analyses. *J Clin Epidemiol* 2016;**79**:70–5.
23. **Hernán MA,** Hernández-Díaz S. Beyond the intention-to-treat in comparative effectiveness research. *Clinical Trials* 2012;**9**:48–55.
24. **Hernán MA,** Robins JM. Per-protocol analyses of pragmatic trials. *N Engl J Med Overseas Ed* 2017;**377**:1391–8.
25. **Toh S,** Hernán MA. Causal inference from longitudinal studies with baseline randomization. *Int J Biostat* 2008;**4**:1557–4679.
26. **Hernán MA,** Hernández-Díaz S, Robins JM. Randomized trials analyzed as observational studies. *Ann Intern Med* 2013;**159**:560–2.
27. **Krieger N,** Davey Smith G. The tale wagged by the DAG: broadening the scope of causal inference and explanation for epidemiology. *Int J Epidemiol* 2016;**45**:1787–808.
28. **Hernán MA.** Does water kill? A call for less casual causal inferences. *Ann Epidemiol* 2016;**26**:674–80.
29. **Hume D.** *An Enquiry Concerning Human Understanding*. Cambridge: Cambridge University Press, 1748. (reprinted 2007).