Causal Inference and Non-randomized Experiments

Michail Katsoulis, Nandita Mitra and A. Floriaan Schmidt

Abstract

Traditionally, machine learning and artificial intelligence focus on problems of diagnosis or prognosis. Answering questions on whether a patient might have a certain disease (diagnosis) or is at risk of future disease (prognosis). In addition to these problems, one might be interested in identifying causal factors which can provide information on how to *change* disease onset or disease progression. In this chapter we introduce the potential outcomes framework, which provides a structured way of conceptualizing questions on causality. Using this framework we discuss how randomized and non-randomized experiments can be conducted, and analyzed, to obtain estimates of the likely causal effect an exposure may have on an outcome.

Keywords

Potential outcomes framework · Non-randomized study · Randomized controlled trials · G-formula · Inverse probability weighting

1 Causal Effects and Potential Outcomes

Researchers often conclude that a factor X is *associated* (or *correlated*) with an outcome Y. However, it may be of interest to be able to conclude that factor X *causes* outcome Y. Causal inference methods aim to answer questions such as: Do Covid masking restrictions reduce coronavirus rates? Does chemotherapy plus radiotherapy increase survival in women with endometrial cancer? Does physical therapy prevent back pain after surgery? Commonly used analytic designs and approaches may only allow one to conclude that these interventions are merely associated with the outcomes. For example, say a study concludes that prostatectomy (surgery to remove the prostate) is *associated* with increased survival among men over the age of 65 with stage III prostate cancer. One interpretation would be that elderly men who received a prostatectomy tended

M. Katsoulis
MRC Unit for Lifelong Health and Ageing, University College London,London, UK

N. Mitra
Division of Biostatistics, University of Pennsylvania, Philadelphia, USA

A. F. Schmidt (✉)
Department of Cardiology; Amsterdam University Medical Centres, Amsterdam, The Netherlands
e-mail: a.f.schmidt@amsterdamumc.nl

Institute of Cardiovascular Science; University College London, London, UK

Division of Heart and Lungs, University Medical Center Utrecht, Utrecht, Netherlands

F. W. Asselbergs et al. (eds.), *Clinical Applications of Artificial Intelligence in Real-World Data*,
https://doi.org/10.1007/978-3-031-36678-9_7

to have longer survival compared to elderly men who did not receive a prostatectomy. On the other hand, say a study concludes that tacrolimus (a skin ointment) *causes* a reduction in skin inflammation in patients with atopic dermatitis. A possible interpretation here would be that tacrolimus, if hypothetically applied to the entire patient population, results in a lower overall skin inflammation rate in this patient population as compared to the hypothetical setting in which no tacrolimus was administered. In the former example, we are making a comparison of outcomes on the basis of treatment actually received. In the latter example, we are making a comparison of two hypothetical scenarios, i.e., the entire population either taking or not taking the treatment. The latter example is what is called a *causal effect* and is the focus of the field of causal inference [1, 2].

Of note, whether association or causation is of importance is fully dependent on the research question at hand. For instance, in cardiovascular research, there is an interest in investigating gender differences in the occurrence of cardiovascular disease. This may have a partial causal explanation or may reflect historical and societal disparities in cardiovascular care between genders. Regardless, having knowledge on the association of gender and disease outcomes can help with clinical aspects of preventive care, diagnosis, and prognosis irrespective of causality. Many researchers feel that causal claims can only be made when the exposure of interest can be *intervened* upon (e.g. dosage of a medication) rather than inherent characteristics such as race or gender. For example, there is an ongoing discussion on whether one can consider race to be a cause since it is not manipulable [3].

Formal causal theory and methods are needed in order to obtain a causal interpretation. Let's first consider a simple linear regression model:

$$Y = \beta_0 + \beta_1 X + \varepsilon;\ \varepsilon \overset{iid}{\sim} N(0, \sigma^2)$$

In this model, we often interpret β_1 by saying "a one unit increase in X is expected to lead to an increase in Y of β_1 units". In reality, we simply observe some people with $X = x$ and other people with $X = x'$. Often we do not observe a change from x to x' in any single person. This then leads to the problem of how to infer causality. In order to define causal effects of interest there are two important components we must specify: (1) a model for the observed data and (2) causal assumptions (which we define in the next section). Causal assumptions are the link between our observed data and causal effects of interest; however, they are often not verifiable.

Here, we introduce the potential outcomes (counterfactual) framework first described by Rubin [4, 5] in order to aid in defining causal effects. We start with common notation. First, let A denote intervention. This can be defined as anything from a medical treatment, policy intervention, or exposure. Note that capital A is a random variable and lowercase a refers to a particular realization of the random variable A. For example, we can say $A = 1$ if a flu vaccine is received and $A = 0$ otherwise. A_i refers to the treatment status of subject i. Next, we let Y denote an outcome of interest which could be continuous (e.g. cholesterol levels), discrete (e.g. cancer remission or not), time to event (e.g. survival), or multidimensional (e.g. longitudinal measures of a biomarker). For example, we can say $Y = 1$ if you experience a recurrence of breast cancer within 5 years and $Y = 0$ otherwise.

We can think of potential outcomes as the outcomes we *would* see under each possible treatment option. For now, we consider the simplest scenario where treatments take place at one point in time; later in the chapter we address treatments over time. Here, Y^a is the outcome that would be observed if treatment was set to $A = a$. Each person has potential outcomes $\{Y^a; a \in \mathcal{A}\}$. For instance when the treatment is binary, Y^0 is the outcome if treated and Y^1 is the outcome if not treated.

Let's look at an example where the outcome is time to event. If treatment is influenza vaccine and the outcome is the time until the individual gets the flu, we would use the following notation:

Y^1: time until the individual would get the flu if they received the flu vaccine,

Y^0: time until the individual would get the flu if they did not receive the flu vaccine.

A second example, where the outcome is binary, is as follows. If treatment is local (A = 1) versus general (A = 0) anesthesia for hip fracture surgery and the outcome (Y) is major pulmonary complications we would use the notation:

Y^1: equal to 1 if major pulmonary complications and equal to 0 otherwise, if given local anesthesia,
Y^0: equal to 1 if major pulmonary complications and equal to 0 otherwise, if given general anesthesia.

Now, the *observed* outcome Y is the outcome under the treatment that a subject actually receives; that is, $Y = Y^A$. In most studies, where participants receive either an intervention treatment or a comparator treatment, for a single subject one can only observe Y^1 or Y^0, and the outcome under the complimentary treatment can be thought of as missing. Counterfactual outcomes are ones that would have been observed had the treatment been different. If a person's treatment was $A = 1$, then their counterfactual outcome is Y^0. If that person's treatment was $A = 0$, then their counterfactual outcome is Y^1.

Let's look at the influenza example again to understand counterfactual outcomes. The causal question we ask is "Did influenza vaccine prevent me from getting the flu?". What actually happened:

1. I got the vaccine and did not get sick.
2. My actual exposure was $A = 1$.
3. My observed outcome was $Y = Y^1$.

What would have happened (contrary to fact) had I not gotten the vaccine? Would I have gotten sick?

1. My counterfactual exposure is $A = 0$.
2. My counterfactual outcome is Y^0.

Before the treatment decision is made, any outcome is a potential outcome: Y^0 and Y^1. After the study, there is an observed outcome, $Y = Y^A$, and counterfactual outcomes Y^{1-A}. Counterfactual outcomes Y^0, Y^1 are typically assumed to be the same as potential outcomes Y^0, Y^1. Thus, these terms are often used interchangeably.

Note that so far we have implicitly assumed that the treatment given to one subject does not affect the outcome for another subject, i.e., $Y_i^{a_i,a_j} = Y_i^{a_i,a_j'}$. In other words, they are independent. If this assumption holds, we can simply write the potential outcome for subject i as only dependent on a_i (one index). However, in many situations, this assumption could be violated such as in the setting of infectious disease. For instance, vaccinating one person in a household might reduce risk of disease among others in the household. This is known as interference.

Now that we have defined potential outcomes, we can formally define causal effects. In general, we say that A has a causal effect on Y if Y^1 differs from Y^0. For example, let's say A is whether or not you take a cold medication (A = 1 you take it, A = 0 you don't) and Y is that your sore throat goes away after an hour (Y = 1 it goes away, Y = 0 it doesn't). Clearly, the statement "I took the cold medicine and my sore throat is gone, therefore the medicine worked" is not proper causal reasoning. This claim is equivalent to $Y^1 = 1$. But what would have happened had you not taken the medicine ($Y^0 =$)? There is only a causal effect if $Y^1 \neq Y^0$. This bring us to the "fundamental problem of causal inference" which stems from the issue that we can only observe one potential outcome for each person. However, with certain assumptions, we can estimate population level (average) causal effect which we will focus on next. In other words, it is possible to answer: What would the rate of sore throat cure be if everyone took the cold medicine versus if no one did? However, without very strong assumptions, we cannot identify individual causal effects that would allow us to answer: What would have happened to me if I had not taken the cold medicine?

Let's first consider individual causal effects. Consider a simple case of binary treatment ($A =$ 1 if treated) and a binary outcome ($Y = 1$ if died). There are four types of individual causal effects [6].

Causal type	Y^0	Y^1	$\delta = Y^1 - Y^0$
Treatment fatal	0	1	1
Always live	0	0	0
Always die	1	1	0
Treatment curative	1	0	−1

Now, let's suppose we have a randomized study (A is randomized) with n participants and there is perfect compliance (all of the study participants adhere to the treatment they are randomized to). In this study, we never observe Y^0 *and* Y^1 for any individual. Instead, we have a random sample of Y^1's and a random sample of Y^0's. We cannot identify δ for any individual. However, we can identify the marginal probabilities $\mathbb{P}(Y^1 = 1)$ and $\mathbb{P}(Y^0 = 1)$. Importantly, We can also identify $\mathbb{E}(\delta)$.

Consider an example where we know that $\mathbb{P}(Y^1 = 1) = 0.1$ and $\mathbb{P}(Y^0 = 1) = 0.2$. In this example, the treatment reduces risk on average by 0.1. We can first write out these marginal probabilities in terms of joint probabilities:

$$\mathbb{P}(Y^1 = 1) = \mathbb{P}(Y^1 = 1, Y^0 = 1) + \mathbb{P}(Y^1 = 1, Y^0 = 0),$$
$$\mathbb{P}(Y^0 = 1) = \mathbb{P}(Y^1 = 1, Y^0 = 1) + \mathbb{P}(Y^1 = 0, Y^0 = 1).$$

We can then write out three examples of the potential outcomes distributions that are consistent with the observed data as follows:

Causal type	Ex1	Ex2	Ex3
Treatment fatal	0	0.05	0.1
Always live	0.8	0.75	0.7
Always die	0.1	0.05	0
Treatment curative	0.1	0.15	0.2

So, for instance, in Ex 1:

$$\begin{aligned}\mathbb{P}(Y^1 = 1) &= \mathbb{P}(Y^1 = 1, Y^0 = 1) \\ &\quad + \mathbb{P}(Y^1 = 1, Y^0 = 0) \\ &= 0.1(\text{always die}) \\ &\quad + 0(\text{treatment fatal}) = 0.1,\end{aligned}$$
$$\begin{aligned}\mathbb{P}(Y^0 = 1) &= \mathbb{P}(Y^1 = 1, Y^0 = 1) \\ &\quad + \mathbb{P}(Y^1 = 0, Y^0 = 1) \\ &= 0.1(\text{always die}) \\ &\quad + 0.10\,(\text{treatment is curative}) = 0.2.\end{aligned}$$

The average causal effect (ACE) is one of the most common causal targets of inference used to compare treatments/exposures. The ACE is given by $\mathbb{E}(Y^1 - Y^0)$. This is the average outcome if everyone had been treated versus if no one had been treated; Fig. 1. Importantly, this is typically not equal to $\mathbb{E}(Y|A = 1) - \mathbb{E}(Y|A = 0)$ which is the average outcome in those who were treated versus the average outcome in those who were not treated; Fig. 2. Specifically, in non-randomized studies, patients who receive a treatment (say surgery) may be very different than those who do not. For instance, those who are deemed fit to withstand surgery may be younger, more healthy, and are less likely to smoke than those who are chosen not to receive surgery.

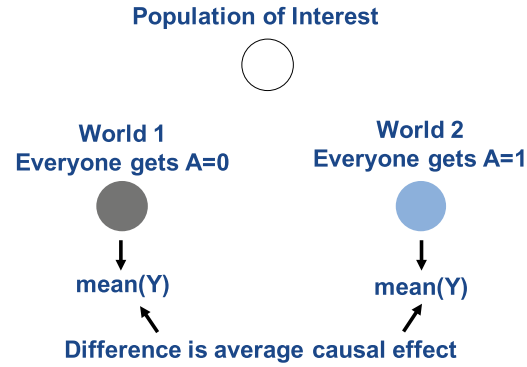

Fig. 1 The average causal effect

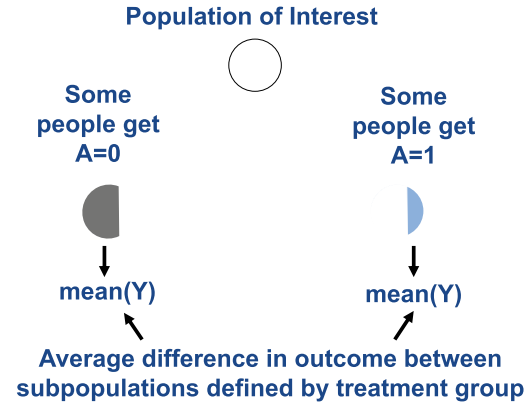

Fig. 2 Effect of a treatment in the real world

In addition to the ACE, $\mathbb{E}(Y^1 - Y^0)$, other causal estimands of interest may include the causal risk ratio, $\mathbb{E}(Y^1)/\mathbb{E}(Y^0)$, the average causal effect among a subgroup defined by V, $\mathbb{E}(Y^1 - Y^0|V = v)$, and the average treatment effect among the treated (ATT) given by $\mathbb{E}(Y^1 - Y^0|A = 1)$ [2]. The ATT, for instance, is a useful estimand when there is interest in the effect of an intervention (say, a treatment of hypertension) on those who received the intervention.

2 Necessary Conditions for Causality

2.1 Randomized Studies with Perfect Compliance

In Sect. 1, we formulated causal effects in terms of potential outcomes. Since potential outcomes are not fully observed we need to make some assumptions in order to be able to estimate (or identify) causal estimands of interest from the observed data. These are called identifying assumptions. Let's first consider a randomized study where there is perfect compliance. In other words, if R is the randomization indicator and A is an indicator of the treatment that is actually taken, then if there is perfect compliance in the trial, $R = A$. We will again consider the potential outcomes Y^0 and Y^1. In randomized trial with full compliance, clearly R is independent of the potential outcomes Y^0 and Y^1. We can express this independence in two different ways using the concepts of *ignorability* and *exchangeability* [6].

Ignorability is stated as $\mathbb{P}(R = 1|Y^0, Y^1) = \mathbb{P}(R = 1)$. In other words, treatment assignment does not depend on the potential outcomes. Say treatment assignment depends on the flip of a coin. Clearly the flip of the coin does not depend on the potential outcomes. Now, if everyone has some non-zero chance of being randomized to the treatment arm, we achieve *strong ignorability*. This assumption that $0 < \mathbb{P}(R = 1) < 1$ is called *positivity* and in this case refers to the fact that we have experimental treatment assignment. Another way to express independence is the concept of *exchangeability*. We can state exchangeability as $f(Y^0, Y^1|R = 1) = f(Y^0, Y^1|R = 0) = f(Y^0, Y^1)$ (where f is the distribution of the potential outcomes). In other words, subjects randomized to $R = 1$ or $R = 0$ are representative of all subjects with respect to the potential outcomes. They are exchangeable.

Exchangeability implies that $f(Y^1) = f(Y^1|R = 1) = f(Y|R = 1)$ and $f(Y^0) = f(Y^0|R = 0) = f(Y|R = 0)$. What we mean by this is that in a randomized trial with perfect compliance, the observed data (the observed outcome Y and the randomization indicator R) are enough to identify the distributions of the potential outcomes, allowing us to estimate causal effects.

Often exchangeability is denoted simply as $Y^a \amalg A$, which can be generalized to include conditional exchangeability $Y^a \amalg A|L$, for covariate L.

2.2 Observational Studies

Randomization allows us to assume, on average, that subjects in different treatment arms are similar to each other on all important factors, whether those factors are measured or not; see Sect. 3. In observational studies, the treatment, intervention or exposure is not controlled by the investigator and by definition is not randomized; although quasi-experiments may naturally occur [7]. Hence, subjects in the treatment group may look very different from those in the comparison group. For instance, men receiving surgery for prostrate cancer may be younger, more likely to be a nonsmoker, and have fewer comorbidities than men who do not receive surgery. The decision, made between the patient and physician, may be based in part by how well the patient is expected to tolerate the surgery. Without accounting for these differences in patient characteristics, the surgery group's survival after surgery may look better than the control group's merely because they were healthier to begin with. As mentioned before, factors that affect both the treatment decision and the outcome are called *confounders*.

Confounding is an important issue that must be addressed in the causal analysis of observational studies. Note that there may also be confounding in randomized trials where there is noncompliance (i.e., $R \neq A$) due to the fact that patients who do not comply with their treatment assignment maybe be different than those who stay on their assigned treatment and those factors may be related to their outcome. This is why RCTs typically do not directly assess treatment effects, but instead estimate the "Intention to Treat" effect; see Sect. 3. If confounders are measured, without meaningful error, we can use standard adjustment methods to control for confounding such as stratification on the confounder, regression adjustment or propensity score methods. Let L be a set of baseline (pre-treatment) covariates. Ignorability in this context means that there is no unmeasured confounding. In other words, if we condition on L, we can control for confounding (there's no hidden bias). If there is no unmeasured confounding, then if we, say, stratify on these covariates, within those strata, we would essentially have a randomized trial. Hence, ignorability can be thought of as conditional randomization where A is independent of the potential outcomes (Y^0, Y^1) given L.

Let's consider an example where treatment assignment depends on the potential outcomes where sicker patients are more likely to be treated. Hence, treated patients have a higher risk of a bad outcome. We need to account for these pre-treatment differences in health. Suppose L are measures of health such as family history of disease, age, weight, smoking status, alcohol, comorbidities, etc. Then within levels of L (i.e., people of the same age, with same co-morbid conditions, of same weight, with same smoking status, etc.), we hope that less healthy patients are not more likely to get treatment. This is the ignorability assumption.

The ignorability setting is comprised of the following three causal assumptions:

- (Condtional) *exchangeability*: treatment is as if randomized conditional on covariates (e.g. within covariate strata).
- *Positivity*: treatment is not assigned in a deterministic fashion (all subjects have a non-zero probability of being assigned to treatment regardless of their covariates). This can be violated when certain treatments are simply unavailable. For example, depending on the urgency, general anesthesia may be the only option available for women undergoing Cesarean section.
- *Consistency*: the potential outcomes are uniquely defined by each subject's own treatment level. This can be violated in situations such as a vaccine trial where one subject's vaccination status can affect another subject's potential outcomes. Other examples include *poorly* defined exposures such changes in BMI which may be occur due to causes such as diet, physical activity or disease.

These identifying assumptions allow us to estimate causal effects directly from the observed data Y, A, L.

3 Randomized Controlled Trials and Estimands of Treatment Effect

In the preceding sections we established a formal definition of causality, and discussed the necessary conditions to interpret a measure of association as an *estimate* of a causal effect.

Historically, discussions on causality have focused on choices in study design, or experiments, where randomized controlled trials (RCTs) remain the unequivocal paradigm. The developed mathematical framework allows for a more detailed discussion of why RCTs provide such a robust design to assess causality. Developing the necessary algebra to describe trial inference is important because it allows us to consider what additional step (analytical or design wise) are required to explore causality in non-randomized (i.e., observational) study designs. Before discussing these analytical methods, we will therefore first further introduce RCTs and touch upon some of the different estimands used in practice (i.e., the type of effect one attempts to estimate).

3.1 Why Association Does Not Imply Causation

Some key features of RCTs include (1) the presence of contemporary intervention and control groups, (2) random allocation of subjects to these groups, and (3) blinding of participants (and often the treating medical professionals) to the group allocations.

If we strip away these three features we are left with a single arm study of subjects who received an intervention. For example the left-panel in Fig. 3 illustrates the results of a hypothetical study assessing the concentration of low-density lipoprotein cholsterol (LDL-C) before ($T = 1$) and after ($T = 1$) subjects were offered treatment with PCSK9 monoclonal antibodies (mAb, a lipid lowering drug [8]). A single arm study would exclusively consider the treated group ($A = 1$). In contrast a "parallel group" design would also consider measurements in participants who did not (decide to) receive treatment ($A = 0$).

An obvious aim would be to attempt to quantify by how much *taking* PCSK9 mAb decreases LDL-C concentrations compared to *not taking* PCSK9 mAb over the same period of time. A relevant estimand would be the *average causal effect*: $\mathbb{E}(Y^0 - Y^1) = \alpha$.

A naive estimate of the treatment lowering effect of PCSK9 mAb would be to use the single arm study and simply take the difference in post- and pre-treatment LDL-C concentrations: $\mathbb{E}(Y|A = 1, T = 1) - \mathbb{E}(Y|A = 1, T = 0)$.
Given that this is a hypothetical example we can also look at the otherwise unknown counterfactual pre- and post-treatment LDL-C concentrations, to clearly see that $\mathbb{E}(Y|A = 1, T = 1) - \mathbb{E}(Y|A = 1, T = 0) \neq \mathbb{E}(Y^0 - Y^1)$. Here we reiterate that by the exchangability assumption $\mathbb{E}(Y^0 - Y^1) = \mathbb{E}(Y^{1,t=1} - Y^{1,t=0})$, meaning that under exchangbility T is ignorable.

As is clear from Fig. 3 the difference in pre- and post-treatment LDL-C concentrations (in treated subjects) does not match the counterfactual difference. In practice this can be caused by a myriad of reasons, often closely linked to the study design and participant sample. In general, one would expect post-treatment concentrations to decrease whenever treatment initiation is (partially) based on a biomarker measurement being elevated (e.g., hypercholesterolemia). Measurements are always subject to (small) random fluctuations, as such the high value necessary to initiate treatment most likely reflects a degree of random upwards variation, which is unlikely to be of the same magnitude in subsequent measurements, hence resulting in a decrease. This well known phenomenon is often referred to as "regression to the mean" [13]. Furthermore, depending on the diagnosis it is not uncommon for a clinician to initiate multiple interventions at the same time. In our example, typically a prescription of lipid lowering therapy would coincide with (referral for) life-style counseling. Similarly, the simple act of prescribing a drug, will incentivse some patients to self-initiate life-style changes (e.g., start exercising more) which will (on average) decrease LDL-C independent of any effect of PCSK9 mAb.

Clearly a single arm study, comparing pre- and post-treatment LDL-C concentrations, will unlikely provide a good estimate of the causal effect of PCSK9 mAb lowering. Instead we could consider conducting a *cohort* study of contemporary participants initiating PCSK9 mAb (the treatment group), compared to a control group of participants who do not receive any treatment; Fig. 3. Assuming for the moment that the control group participants were "blinded" from the fact they did not receive any treatment, the difference in LDL-C concentration of the control group participants is identical to that of the counterfactual (i.e, comparing measurements at $T = 0$ to $T = 1$). However, because treatment was not initiated at random, we see that the control group measurements are substantially lower than that of the counterfactual; simply reflecting that medical professionals treat patients at risk. As such, despite having a control group, the difference between the treatment and control group will not equal our inferential target.

3.2 Treatment Estimands in Trials

While by itself inclusion of a control group does not typically result in a causal effect estimate of our inferential target $\mathbb{E}(Y^0 - Y^1) = \alpha$, it

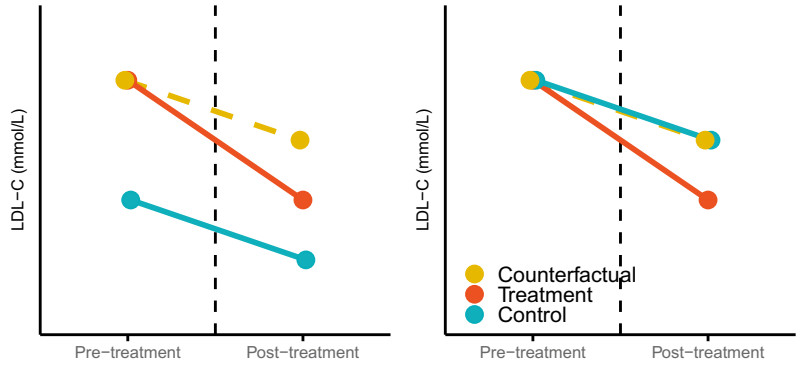

Fig. 3 Causal contrasts in a study evaluating changes in LDL-C concentration. The left-panel represents a possible *non-randomized* study, and the right-panel a possible scenario for a *randomized* study. Notice that the x-axis values are slightly dodged to help identify overlapping points and lines

does provide a suggestion how we could further improve our study – we could randomize treatment assignment! The right-panel of Fig. 3 illustrates this, showing agreement between the control group measurements and the counterfactual LDL-C measurements. In this setting we will have that $\mathbb{E}(Y|A=1,T=1)-\mathbb{E}(Y|A=0,T=1)=\mathbb{E}(Y^0-Y^1)$, implying that stringently designed RCTs provide relevant causal estimates.

If we simply focus on time $T=1$ the above estimator $\mathbb{E}(Y|A=1)-\mathbb{E}(Y|A=0)$ is often referred to as the "as-treated" (AT) estimator. Interestingly, and contrary to the above derivations, the AT estimator is considered to be a biased estimator. To see why, we will move a way from the hypothetical trial with perfect compliance (see Sect. 2.1), and expand our example to differentiate between treatment allocation Z, and the actual treatment taken A. To illustrate the difference, note that adherence is defined as

$$\mathbb{P}(A=1|Z=1)-\mathbb{P}(A=1|Z=0)=\phi$$

where values close to 1 indicate subjects generally took the allocated treatment, and smaller values indicate non-adherence to treatment allocation.

In the previous subsection we thus made the implicit and unrealistic, assumption of complete adherence. Worse, as shown in Fig. 4, in the presence of non-adherence the association between A and Y will be subject to confounding by common cause(s) L, violating the exchangeability assumption: $Y^a \not\perp\!\!\!\perp A$. Hence, in the presence of non-adherence, the AT-estimator will never equal the true causal treatment effect unless we are willing to assume there are no L at all. We could of course decide to condition on L and create a conditional AT-estimator, however knowledge of L is typically incomplete and above all it would be difficult to determine when such conditioning sufficiently addressed confounding - defeating the purpose of a trial: balancing on known *as well as* unknown confounders.

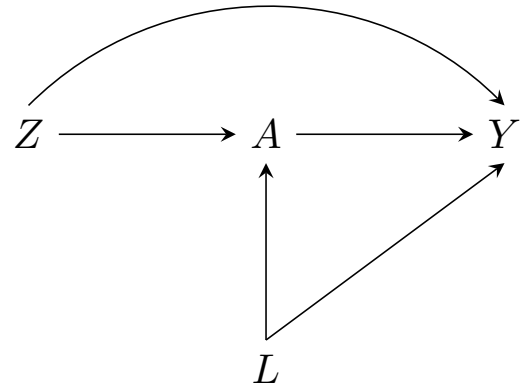

Fig. 4 A directed acyclic graph representation of a randomized control trial. Here Z represent treatment allocation, X treatment itself, Y the primary outcome, L measured and unmeasured common causes of X and Y. The directed paths (i.e., arrows) represents a cause and effect relation of unspecified magnitude which may also include zero (i.e., when there is no path)

Due to the frailty of associating A with Y, trials commonly forgo this estimand entirely and perform an "intention to treat" (ITT) analysis, with estimator

$$\mathbb{E}(Y|Z=1)-\mathbb{E}(Y|Z=0)=\alpha\phi+\tau.$$

Here α is the effect treatment allocation has on the outcome mediated through A. Additionally τ represents the possibility that treatment allocation may affect the outcome indirectly, sidestepping A. For example, subjects allocated to the untreated group may decide to exercise more. Inclusion of $\tau \neq 0$ is of course problematic because the ITT estimator no longer solely evaluates effects mediated through A, and a trial may incorrectly suggest treatment is beneficial.

By defining the ITT estimator as the sum of the true causal treatment effect (α) multiplied by adherence (ϕ) and the direct effect (τ) of treatment allocation, we can finally comment on the relevance of blinding in trial design. Blinding trial participant and staff, to knowledge of the allocated treatment ensures that, on average, enrolled subjects behave the same-way irrespective of Z, and thus that we can assume $\tau = 0$. The results of this is that $\mathbb{E}(Y|Z=1) - \mathbb{E}(Y|Z=0) \neq 0$ implies that $\alpha \neq 0$, irrespective of treatment adherence. In many ways randomization and blinding are complementary strategies to ensure participant groups are (on average) similar at baseline (*randomization*) and behave similar during follow-up (*blinding*).

Assuming blinding and randomization were conducted adequately the ITT estimator thus equals α only if participants completely adhered to treatment allocations. In all other settings the ITT estimator is a biased estimator of the causal treatment effect and will not equal α. The ITT estimator is thus a flawed *estimator*. Nevertheless, it does have desirable properties, 1) when sufficiently blinded the ITT estimator will (on average) be zero whenever $\alpha = 0$, and therefore 2) it often provides a robust indicator of effect direction (i.e., whether treatment is beneficial or not).

While the ITT estimator does not in general provide an estimate of our inferential target α, we can however use it to perform an instrumental variable (IV) analysis, which assuming $\tau = 0$, will on average equal our inferential target:

$$\begin{aligned}\frac{\mathbb{E}(Y|Z=1) - \mathbb{E}(Y|Z=0)}{\mathbb{P}(X=1|Z=1) - \mathbb{P}(X=1|Z=0)} &= \frac{\alpha\phi}{\phi}, \\ &= \alpha.\end{aligned}$$

This IV estimator essentially corrects the ITT estimate for the amount of non-adherence, and in doing so obtains an estimate of $\alpha = \mathbb{E}(Y^1 - Y^0)$. Of course all this is under the assumption the trial has been appropriately randomized and blinded, which we can elegantly frame as ignorabililty. It is important to reiterate that the ignorabililty assumption refers to the randomized groups and as such all the previously discussed estimands will not generally hold for individuals, and do not represent *individual* causal effects unless there are convincing reasons to expect an absence of between-patient treatment heterogeneity [9]. Note Schmidt et al. 2018 [10] discusses IV analysis in the a setting of a meta-analysis of potentially unblinded trials, where $\tau \neq 0$.

4 Non-randomized Experiments of Time-Fixed Exposure and Confounders

As discussed RCTs are the gold standard to explore causal were design steps such as randomization and blinding are essential to ensure the three critical assumptions (exchangeability, positivity and consistency) are likely true. In many cases one may not be able to perform a RCT, for example an RCT may be prohibitively costly, or patients may be difficult to recruit. Moreover, randomisation may not always be ethical, for example when the comparator intervention can cause harm (e.g., shame surgeries). Because of these reasons only a small proportion (15–20%) clinical practice guidelines are based on an 'A' level of evidence (based on multiple RCTs), and most rely on evidence from non-randomized (observational) studies [11, 12]. It is therefore essential to be able to identify, and conduct, high quality analyses using non-randomised study designs.

Non-randomised studies, in contrast to RCTs, may be much less convincing to assess causal inferences for treatments/interventions. As an

example, take an observational study from electronic health records where a researchers is interested in evaluating the effect statin prescription may elicit on the incidence of cardiovascular disease. Those who initiated statins are more likely to be in a worse health state compared to those who did not initiate statins. In other words, it is very likely that we have problems due to confounding by indication. If we have sufficiently detailed information from for example EHR capturing all the confounding variables, then there are many options to account for such confounding bias; otherwise, our analysis will suffer from unmeasured confounding. In the next paragraphs, we will explain in detail how to deal with non-randomised experiments of time-invariant exposures.

Let's focus on the following example: in the Table below, we have 12 patients with measured data on statin initiation X (0 = untreated, 1 = treated) and whether they developed cancer after 10 years, i.e. cancer incidence Y (0 = no cancer, 1 = cancer). The question of interest is: What is the effect of statin initiation on cancer incidence?

Participant	Other comorbidities L	Statin X	Cancer Y
Isabella	0	0	0
Oliver	0	0	1
Rachel	0	0	0
George	0	0	1
Rebecca	0	1	0
Oscar	0	1	1
Natalie	1	0	1
Tom	1	0	0
Margaret	1	0	0
Charles	1	1	1
Olivia	1	1	0
Harry	1	1	0

From these data, we observe that statin initiation (X) is associated with cancer incidence (Y): the probability of developing cancer among those who initiated statin therapy is $\mathbb{P}(Y = 1|X = 1) = 2/5 = 0.40$, while the probability of developing cancer among those who did not initiated statin therapy is $\mathbb{P}(Y = 1|X = 0) = 3/7 = 0.43$. The observed risk difference, is $\mathbb{P}(Y = 1|X = 1) - \mathbb{P}(Y = 1|X = 0) = -1/35$. At face value the observed difference in cancer incidence between statin initiators might be taken to imply statin prescription is carcinogenic. Depending on the plausibility of the in sect. 2.2 assumptions, the observed difference may be distinct from our inferential target estimand $\mathbb{P}\left(Y^{X=1} = 1\right) - \mathbb{P}\left(Y^{X=0} = 1\right)$

Let's assume that, in an over-simplistic scenario, the two groups have the same characteristics (age, sex socioeconomic status, family history of cancer etc.), apart from other comorbidities L. In other words, if we account for other comorbidities appropriately in this sample, we will emulate randomisation successfully.

Under the consistency, (conditional) exchangeability, and positivity assumptions (see sect. 2.2), we can estimate $\mathbb{P}\left(Y^{X=1} = 1\right) - \mathbb{P}\left(Y^{X=0} = 1\right)$ accounting for L, using standard regression modelling, standardisation or inverse probability of weighting.

4.1 Analytical Methods to Estimate the Effect of Time-Fixed Exposures

4.1.1 Regression Modelling

Standard regression modelling, in which we include all the (*likely*) confounders as covariates is a popular way of dealing with time-fixed confounders. In the example presented above, we could choose to create a logistic regression model, given that the outcome is binary. In that case, the estimand of interest would be the causal odds ratio, i.e.

$$\text{causal odds ratio} = \frac{\dfrac{\mathbb{P}\left(Y^{X=1} = 1\right)}{\mathbb{P}\left(Y^{X=1} = 0\right)}}{\dfrac{\mathbb{P}\left(Y^{X=0} = 1\right)}{\mathbb{P}\left(Y^{X=0} = 0\right)}}$$

which will be equal to the observed (conditional) odds ratio

$$\text{observed odds ratio} = \frac{\frac{\mathbb{P}(Y=1|X=1,L=l)}{\mathbb{P}(Y=0|X=1,L=l)}}{\frac{\mathbb{P}(Y=1|X=0,L=l)}{\mathbb{P}(Y=0|X=0,L=l)}}$$

under the assumptions described in sect. 2.2. Moreover, the observed odds ratio for X can be easily calculated from a logistic regression where the outcome is Y and we adjust for L, i.e.

$$\text{logit}(\mathbb{P}(Y=1|X,L)) = a_0 + a_1 X + a_2 L$$

In the example of Fig. 5, the odds ratio OR=e^{a_1} is equal to 1, which means that

$$\mathbb{P}\left(Y^{X=1}=1\right) = \mathbb{P}\left(Y^{X=0}=1\right).$$

4.1.2 Standardisation—G-Formula

The G-formula provide an alternative approach to account for possible confounding. Here we wish to obtain an unbiased estimate of the outcome risk under different interventions X leveraging the fact that conditional on L, the counterfactial outcome is independent of X, e.g. the conditional exchangeability assumption holds: $Y^x \amalg X|L$. Specifically, the observed conditional risk under treatment is equal to the counterfactual risks:

$$\mathbb{P}(Y=1|X=x,L=l) = \mathbb{P}\left(Y^{X=x}=1|L=l\right)$$

To calculate the $\mathbb{P}\left(Y^{X=x}=1\right)$, we will use the formula

$$\mathbb{P}\left(Y^{X=x}=1\right) = \sum_l \mathbb{P}(Y=1|X=x,L=l) \times \mathbb{P}(L=l), l \in \{0,1\}$$

In other words,

$$\mathbb{P}\left(Y^{X=1}=1\right) = \mathbb{P}(Y=1|X=1,L=0)\times\mathbb{P}(L=0) + \mathbb{P}(Y=1|X=1,L=1)\times\mathbb{P}(L=1)$$

and

$$\mathbb{P}\left(Y^{X=0}=1\right) = \mathbb{P}(Y=1|X=0,L=0)\times\mathbb{P}(L=0) + \mathbb{P}(Y=1|X=0,L=1)\times\mathbb{P}(L=1)$$

Risk had all individuals received treatment: $\mathbb{P}\left(Y^{X=1}=1\right)$
We know that the risk if all individuals had been treated is 1/2 in the 6 individuals with $L=0$ and 1/3 in the 6 individuals with $L=1$. Therefore, the risk if all individuals in the population had been treated will be a weighted average of 1/2 and 1/3 in which each group receives a weight proportional to its size. Since 50% of the individuals are in group $L=0$ and 50% of the individuals in $L=1$ The weighted average will be (1/2 × 0.5) + (1/3 × 0.5) = 0.42.

Risk had no individuals received treatment: $\mathbb{P}\left(Y^{X=0}=1\right)$
We know that the risk if all individuals had not been treated is 2/4 in the 6 individuals with $L=0$ and 1/3 in the 6 individuals with $L=1$. Therefore, the risk if all individuals in the population had not been treated will be a weighted average of 1/2 and 1/3 in which each group receives a weight proportional to its size. Since 50% of the individuals are in group $L=0$ and 50% of the individuals in $L=1$. The weighted average will be (2/4 × 0.5) + (1/3 × 0.5) = 0.42.

4.1.3 Inverse Probability Weighting

Inverse probability weighting (IPW) is a further alternative method to account for confounding, here one creates a pseudo-population in which treatment is independent of the covariates L. Treated and the untreated are (unconditionally) exchangeable in the pseudo-population because the X is independent of L. In other words, the arrow from the covariates L to the treatment X is removed (see Fig. 5).

Using IPW, we weight each individual by the inverse of the probability of receiving the treatment (exposure), conditional on the confounders.

$$\text{IPW} = \frac{1}{\mathbb{P}(X|L)}$$

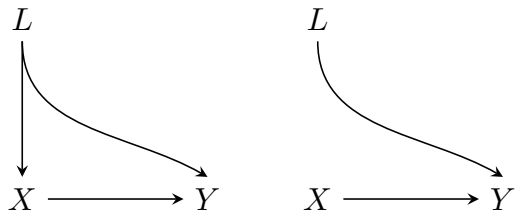

Fig. 5 Directed acyclic graphs in the population (right panel) and the pseudo-population (left panel) creates by inverse probability weights

In our example, the created pseudo-population will be twice as large as the original population (see Fig. 5 in the right). Under conditional exchangeability $Y^x \amalg X|L$ in the original population, treatment is randomized in the pseudo-population i.e. treated and the untreated are (unconditionally) exchangeable in the pseudo-population because the X is independent of L From the pseudo-population, we can calculate $\mathbb{P}\left(Y^{X=1}=1\right)$ and $\mathbb{P}\left(Y^{X=0}=1\right)$.

That is, the associational risk ratio in the pseudo-population is equal to the causal risk ratio in both the pseudo-population and the original population.

In the pseudo-population (see Fig. 6 we observe that a) among the untreated the expected number of cancer events are 5 in 12 individuals, i.e. $\mathbb{P}\left(Y^{X=0}=1\right)=5/12=0.42$, and b) among the treated the expected number of cancer events are 5 in 12 individuals, i.e. $\mathbb{P}\left(Y^{X=1}=1\right)=5/12=0.42$. We therefore find that there is no causal effect of treatment X on the outcome Y, i.e., $\mathbb{P}\left(Y^{X=0}=1\right)=\mathbb{P}\left(Y^{X=0}=1\right)$.

5 Non-randomized Experiments of Time-Dependent Exposure and Confounders

In this chapter, we will explain how to deal with non-randomised experiments of time-dependent exposures. We will first explain why standard methods (e.g., outcome regression models) fail to provide correct estimates of average causal exposure effect estimate correctly the causal effect when time-dependent confounders are affected by exposure (treatment) history.

5.1 Why Standard Methods May Fail

In Fig. 7 treatment A can change with time $t \in \{0, 1\}$, as do the confounders L. In this example, L_1 is both a confounder (between A_0 and Y) and a mediator (between A_1 and Y), in other words, we should both adjust for L_1 (because it is a confounder) and not adjust for L_1 (because it's a mediator). If we adjust for L_1, we induce bias because we block part of the effect of A_0 through L_1. However, if we do not adjust for L_1, the estimated effect will be biased through the back door pathway $A_1 \leftarrow L_1 \rightarrow Y$, which induces confounding bias.

5.2 Use of G-Methods to Overcome the Problem

Below, we will present an example we IPW is used account for time-varying confounding without removing exposure effects mediated by L_0 and L_1. IPW creates a pseudo-population in which the arrows headed to A_0 and A_1 do not exist and hence we do not need to adjust for L_0 and L_1 (Fig. 8).

For example, in the table below, if we want to estimate the causal contrast $\mathbb{E}(Y^{\bar{a}=(1,1)}) - \mathbb{E}(Y^{\bar{a}=(0,1)})$,when $\bar{a}$ is the treatment history, then we should estimate the associational risk difference in the pseudo-population $\mathbb{E}(Y|A_0=1, A_1=1) - \mathbb{E}(Y|A_0=0, A_1=1)$ created by the weights

$$\begin{aligned}\text{IPW} &= \frac{1}{\mathbb{P}(A_0|L_0) \times \mathbb{P}(A_1|L_0, A_0, L_1)} \\ &= 282.5 - 281.82 = 0.68.\end{aligned}$$

Please note, that we would not get the correct answer for the causal effect of A on Y if

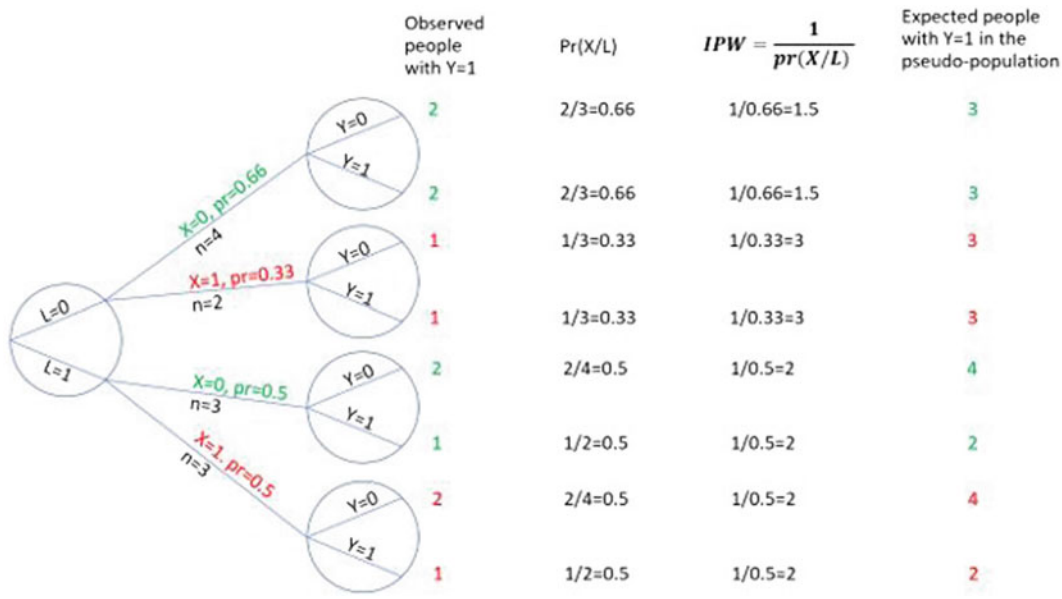

Fig. 6 Calculation of inverse probability weights (IPW)

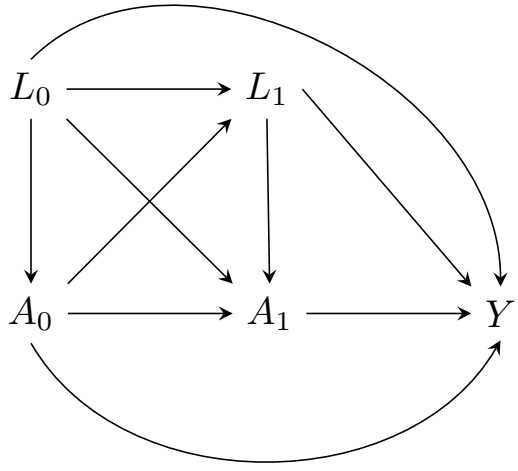

Fig. 7 A directed acyclic graph with time-dependent confounders L affected by treatment history

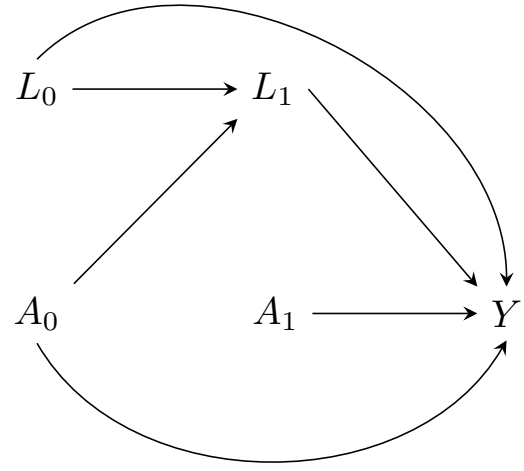

Fig. 8 A directed acyclic graph with time-dependent confounders L affected by treatment history in the pseudo-population, created by IPW

L_0	A_0	L_1	A_1	N	$E[Y \mid A_0, L_1, A_1]$	$P(A_0 \mid L_0)$	$P(A_1 \mid L_1, A_0, L_0)$	IPW*	N_{ps}	$E[Y \mid A_0, A_1]$	$E_{ps}[Y \mid A_0, A_1]$
0	0	0	0	3000	200	2/3	3/4	12/6=2	6000	When A_0=0, A_1=0	
0	0	0	1	1000	250	2/3	1/4	12/2=6	6000	221.67	227.27
0	0	1	0	1000	150	2/3	1/4	12/2=6	6000	When A_0=1, A_1=0	
0	0	1	1	3000	300	2/3	3/4	12/6=2	6000	243	247.5
0	1	0	0	1000	180	1/3	2/3	9/2	4500	When A_0=0, A_1=1	
0	1	0	1	500	280	1/3	1/3	9	4500	288.57	281.82
0	1	1	0	1250	220	1/3	1/2	6	7500	When A_0=1, A_1=1	
0	1	1	1	1250	240	1/3	1/2	6	7500	297	282.5
1	0	0	0	1000	260	1/2	2/5	5	5000	In general $E[Y^{A_0=a_0,A_1=a_1}] = E_{ps}[Y \mid A_0, A_1]$	
1	0	0	1	1500	300	1/2	3/5	10/3	5000		
1	0	1	0	1000	320	1/2	2/5	5	5000		
1	0	1	1	1500	280	1/2	3/5	10/3	5000	And	
1	1	0	0	2000	260	1/2	2/3	6/2=3	6000		
1	1	0	1	1000	280	1/2	1/3	6	6000	$E_{ps}[Y \mid A_0, A_1] \neq E[Y \mid A_0, A_1]$	
1	1	1	0	750	320	1/2	1/4	8	6000		
1	1	1	1	2250	340	1/2	3/4	8/3	6000		

*IPW $= \frac{1}{P(A_0 \mid L_0) * P(A_1 \mid L_1, A_0, L_0,)}$

1. we do not adjust for L_0 and L_1, because the associational risk difference in the actual population is not causal

$$\mathbb{E}(Y|A_0 = 1, A_1 = 1) - \mathbb{E}(Y|A_0 = 0, A_1 = 1) = 297 - 288.57 = 8.43$$

2. we adjust for L_0 and L_1 (e.g. through standardisation), because the standard methods fail in the context of time dependent confounding affected by prior treatment.

For example, within the strata defined by L_0 and L_1, we have that

$$\begin{aligned}
&L_0 = 0, L_1 = 0 : \mathbb{E}(Y|A_0 = 1, A_1 = 1) \\
&\quad - \mathbb{E}(Y|A_0 = 0, A_1 = 1) = 280 - 250 = 30, \\
&L_0 = 0, L_1 = 1 : \mathbb{E}(Y|A_0 = 1, A_1 = 1) \\
&\quad - \mathbb{E}(Y|A_0 = 0.A_1 = 1) = 240 - 300 = -60, \\
&L_0 = 1, L_1 = 0 : \mathbb{E}(Y|A_0 = 1, A_1 = 1) \\
&\quad - \mathbb{E}(Y|A_0 = 0, A_1 = 1) = 280 - 300 = -20, \\
&L_0 = 1, L_1 = 1 : \mathbb{E}(Y|A_0 = 1.A_1 = 1) \\
&\quad - \mathbb{E}(Y|A_0 = 0.A_1 = 1) = 340 - 280 = 60.
\end{aligned}$$

Accounting for L_0 and L_1 (e.g., through regression adjustment) would give us an estimate of

$$\mathbb{E}(Y|A_0 = 1, A_1 = 1) - \mathbb{E}(Y|A_0 = 0, A_1 = 1)$$

which is equal to

$$\begin{aligned}
&30 \times \mathbb{P}(L_0 = 0, L_1 = 0) - 60 \times \mathbb{P}(L_0 = 0, L_1 = 1) \\
&\quad -20 \times \mathbb{P}(L_0 = 1, L_1 = 0) + 60 \times \mathbb{P}(L_0 = 1, L_1 = 1) \\
&= \frac{30 \times 5500}{23000} - \frac{60 \times 5500}{23000} \\
&\quad - \frac{20 \times 6500}{23000} + \frac{60 \times 5500}{23000} = -0.21,
\end{aligned}$$

which does not correspond to the causal risk difference.

We could also derive unbiased estimates when dealing with time-dependent confounders, affected by prior treatment (exposure) using the other g-methods (i.e. g-formula, g-estimation), however this is beyond the scope of this chapter.

References

1. Pearl J. Causality: models, reasoning, and inference. Cambridge University Press; 2009.
2. Hernan M, Robins J. Causal inference: what If. Chapman & Hall; 2020.
3. Kaufman J, Cooper R. Commentary: considerations for use of racial/ethnic classification in etiologic research. Am J Epidemiol. 2001;154(4):291–8.
4. Rubin DB. Estimating causal effects of treatments in randomized and nonrandomized studies. J Educ Psychol. 1974;66(5):688–701.
5. Rubin DB. Causal inference using potential outcomes. J Am Stat Assoc. 2005;100(469):322–31.

6. Greenland S, Robin J. Identifiability, exchangeability, and epidemiological confounding. Int J Epidemiol. 1986;15(3):413–9.
7. Meyer BD. Natural and quasi-experiments in economics. NBER Technical Working Papers 0170, National Bureau of Economic Research, Inc, Dec. 1994.
8. Schmidt AF, Carter JL, Pearce LS, Wilkins JT, Overington JP, Hingorani AD, Casas JP. Pcsk9 monoclonal antibodies for the primary and secondary prevention of cardiovascular disease. Cochrane Database of Syst Rev. 2020;10.
9. Schmidt A, Klungel O, Nielen M, De Boer A, Groenwold R, Hoes A. Tailoring treatments using treatment effect modification. Pharmacoepidemiology and drug safety. 2016;25(4):355–62.
10. Schmidt A, Groenwold R. Adjusting for bias in unblinded randomized controlled trials. Stat Methods Med Res. 2018;27(8):2413–27.
11. Lai AG, Chang WH, Parisinos CA, Katsoulis M, Blackburn RM, Shah AD, Nguyen V, Denaxas S, Davey Smith G, Gaunt TR, et al. An informatics consult approach for generating clinical evidence for treatment decisions. BMC Medical Inf Dec Making. 2021;21(1):1–14.
12. Fanaroff AC, Califf RM, Windecker S, Smith J, Sidney C, Lopes RD. Levels of evidence supporting American college of Cardiology/American Heart Association and European Society of Cardiology Guidelines, 2008–2018. JAMA. 2019;321:1069–80.
13. Bland JM, Altman DG. Statistic notes: regression towards the mean BMJ 1994; 308(6942):1499 https://pubmed.ncbi.nlm.nih.gov/8019287/