6

Causal Inference from Case-Control Studies

Vanessa Didelez

Leibniz Institute for Prevention Research and Epidemiology & University of Bremen

Robin J. Evans

University of Oxford

6.1 Introduction

In this chapter we consider when it is possible to draw causal conclusions from a case-control study. This requires special attention because causal conclusions make stronger claims than ordinary predictions. For instance, it may not be sufficient to note that people with a higher

alcohol intake are at a higher risk of stroke; rather, we may wish to establish that if we are able to reduce people's alcohol intake, this will have the *effect* of reducing the number of instances of stroke. In other words, we wish to establish that certain *interventions* in the level of an exposure or in a lifestyle choice will result in improved health, and to quantify the strength of such effects. The core problems of causal inference are therefore to estimate and predict the effects of interventions. For background reading on causality and causal inference, see textbooks and overviews by Hernán and Robins (2018), Pearl (2009), Spirtes *et al.* (2000), Rubin (1974), and Dawid (2015).

Readers may be more familiar with the famous Bradford Hill (BH) criteria for causation (Hill, 2015): strength, consistency, specificity, temporality, biological gradient, plausibility, coherence, experiment, and analogy. These are important in order to strengthen causal conclusions in the wider context of scientific investigations. Here, we take a different, but not entirely unrelated, approach, placing the emphasis on first formally defining a causal effect as the effect of an intervention (related to BH's criterion that experimental evidence strengthens causal conclusions), and considering what problems we face when estimating such an effect with case-control data. It can then be formally investigated how and when the Bradford Hill criteria are sensible for a given causal model and scientific query at hand, e.g., an explicit causal model might postulate a plausible dose-response relation reflecting the biological gradient as well as the assumed underlying mechanism. For a recent critical revision of BH's criteria, see Ioannidis (2016).

Taking this formal approach to causal inference, a key question is that of *identifiability*: does the available data, at least in principle, allow us to consistently estimate the desired causal quantity? If the answer is 'no' then this is typically due to *structural bias*. Here, 'in principle' means the sample size is large enough to get reliable nonparametric estimates of distributions. The term 'structural bias' refers to fundamental problems of design and available information – such as the difficulty in distinguishing a 'real' causal effect from correlation due to an unmeasured common cause – and not to particular choices of parametric models or sampling error.

The effect of specific interventions can most unambiguously be investigated in randomised controlled trials (RCTs). However, there are many reasons to use observational data and in particular case-control studies in practice, such as because the disease under investigation is very rare. In case-control studies, we face the following potential sources of structural bias regarding causal inference:

(1) Case-control studies are necessarily observational, so *confounding* is likely to be present.

(2) Case-control studies are retrospective with sampling being conditional on disease status which means there is also a threat of *selection bias* (Hernán *et al.*, 2004).

(3) A consequence of the retrospective sampling is that methods which depend on, or are sensitive to, the marginal distribution of the outcome cannot be used without some modification, since the required information is not generally available. This is potentially relevant to methods of adjusting for confounding as well as to the identifiability of typical causal effect measures, such as the average causal effect (ACE). An effect measure that is not sensitive to the retrospective sampling is the odds ratio (OR) but its interpretation can be complicated by its noncollapsibility (Greenland *et al.*, 1999).

While confounding is a problem of any observational study and has been widely addressed in the causal inference literature (Hernán and Robins, 2018; Pearl, 1995), points (2) and (3) are more specific to case-control studies and will be the focus of this chapter. Note that retrospective sampling gives rise to further sources of bias, such as recall bias or reverse causation which combines measurement error with potential selection bias as addressed later. We can relate our approach to BH's criteria: 'specificity' is partly about excluding sources of bias, and so ensuring that there is no other likely explanation for an

FIGURE 6.1
Directed acyclic graphs representing sampling under (a) an unmatched case-control study; (b) a matched case control study with matching variables C.

association; respecting 'temporality' is especially relevant with case-control data, as seen in the context of recall bias. Note that biases can to some extent be avoided by careful design of a case-control study; in this context we refer the reader to Chapter 2 of this book which describes several case-control design options, outlines classical approaches to analysis, and mentions threats to internal validity. One topic not addressed in the present chapter is causal search, where the aim is to automatically identify a causal structure from case-control data, but see Cooper (1995) and Borboudakis and Tsamardinos (2015).

6.1.1 Basic set-up

Throughout D denotes a binary outcome (disease), with $D = 1$ denoting the cases and $D = 0$ the controls. We write X for the exposure and C for a set of covariates, with respective domains $\mathcal{X}$ and $\mathcal{C}$; in general these may be binary, categorical, or continuous, but for simplicity we will largely restrict ourselves to binary X and categorical C. The index $i = 1, \ldots, n$ is used to indicate different units (e.g., D^i, X^i), but usually this is suppressed. For brevity we will often write (for example) $P(D \mid X)$ instead of $P(D = d \mid X = x)$.

6.1.2 Sampling indicator

It will be important to explicitly formalise aspects of the sampling mechanism, so we introduce a binary indicator S, where $S = 1$ means the unit is present in the sample. It follows that data from a case-control study can usually be regarded as a sample from $P(X, C, D \mid S = 1)$, where the sampling proportion $\tilde{p}_S = P(D = 1 \mid S = 1)$ is often known while the population prevalence $\tilde{p} = P(D = 1)$ may or may not be known. We mainly address methods that do not require knowledge of $\tilde{p}$, but see Section 6.5.

6.1.3 Directed acyclic graphs and conditional independence

We use directed acyclic graphs (DAGs) to express the (structural) assumptions we are willing to make about the joint distribution within the *population*, including how a subset of the population is sampled. For example, in Figure 6.1(a) the sampling variable S only depends on the outcome D, whereas in (b) it also depends on measured covariates due to matching factors in C.

The DAG represents conditional independence constraints on the joint probability distribution over the variables shown as nodes in the graph. In particular, any variable should be conditionally independent of its graphical nondescendants given its graphical parents. (For full details and graphical terminology, see the Appendix.) For random variables A, B, Z we

denote the statement "A is conditionally independent of B given Z" by $A \perp\!\!\!\perp B \,|\, Z$ (Dawid, 1979). This means, for instance, that $P(A \,|\, B, Z) = P(A \,|\, Z)$.

For example, in Figure 6.1(a) the variable S only has the parent D, so in the population

$$S \perp\!\!\!\perp (X, C) \,|\, D; \tag{6.1}$$

in particular this means that $P(X, C \,|\, D, S = 1) = P(X, C \,|\, D)$. In contrast, the additional $C \to S$ edge in Figure 6.1(b) means that now we can only assert (from the DAG) that

$$S \perp\!\!\!\perp X \,|\, (C, D). \tag{6.2}$$

The two conditional independence assumptions (6.1) and (6.2), and the fact that the former implies the latter, will be used many times throughout this chapter.

Note that in a DAG, it is the absence of edges that encodes assumptions; the presence of an edge means merely that we allow the *possibility* of a dependence, not that we claim there necessarily is one. Some independencies cannot be seen graphically; for instance it is common to match such that $D \perp\!\!\!\perp C \,|\, S = 1$, but this is not represented by the graph.

When we include S in our DAGs we express assumptions about the whole population, those that are sampled as well as those that are not. This could include independencies that 'disappear' when we condition on $S = 1$, i.e., that cannot be observed in the sampled population. We address this phenomenon of *selection* again in the following section.

If the graph implies the existence of conditional independencies in the distribution corresponding to the observed data (in particular conditional on $\{S = 1\}$) then these can be checked empirically. Those not conditional on $\{S = 1\}$ should approximately hold in the controls-only data in the situation of a case-control study with a very rare disease.

6.2 Basic concepts of causal inference

When we are interested in causality, it is helpful to use notation that distinguishes causal concepts, such as the effect of an intervention, from associational ones. There are several frameworks for making this distinction, including potential outcomes (Rubin, 1974), structural equations (Goldberger, 1972), the do(·)-operator (Pearl, 2009), or a decision theoretic approach (Dawid, 2015) (a hierarchy can be found in Richardson and Robins (2013)); these are further often supplemented by DAGs, or extensions thereof, as a shorthand to illustrate assumptions made (Spirtes *et al.*, 2000; Pearl, 2009). We do not go into the subtle differences between these approaches here (see discussion by Dawid (2000); Didelez and Sheehan (2007b)), but explain the basic ideas that are essentially common to all of them.

6.2.1 Intervention distribution

We use $P(D \,|\, \mathrm{do}(X = x))$ to denote the distribution of D if X is set to $x \in \mathcal{X}$ by an intervention (Pearl, 2009). If a randomised controlled trial can be carried out, then $P(D \,|\, \mathrm{do}(X = x))$ can immediately be estimated from the corresponding arm of the trial where $X = x$. The intervention distribution is different from a conditional distribution: $P(D \,|\, X = x)$ simply denotes the distribution amongst units in which $X = x$ is *observed* to have happened; the two are generally distinct if there are other common causes or *confounders*.

An alternative approach uses *potential outcomes* D_x to denote the disease status when the exposure or treatment X is set to $x \in \mathcal{X}$ by an intervention. In the case of binary X, we have potential outcomes D_1 and D_0; these are sometimes also called *counterfactuals*

because we can always only observe either D_1^i or D_0^i but not both for a given unit i. For the purposes of this chapter we consider $P(D \mid \mathrm{do}(X = x))$ as equal to the distribution of the corresponding potential outcome $P(D_x)$, and focus on quantities that are functions of these distributions.

6.2.2 Causal effect and causal null hypothesis

In analogy to a randomised trial, we say that X *has a causal effect on* D if, for two different exposure values $x \neq x'$, the intervention distributions differ: $P(D \mid \mathrm{do}(X = x)) \neq P(D \mid \mathrm{do}(X = x'))$. Accordingly, we formalise the causal null hypothesis as

$$H_0^{\emptyset} : P(D \mid \mathrm{do}(X = x)) = P(D \mid \mathrm{do}(X = x')) \quad \text{for all } x \neq x'.$$

For simplicity, instead of considering the whole distribution it is common to use causal parameters that contrast particular aspects of the intervention distributions: for instance the difference in means $E(D \mid \mathrm{do}(X = x)) - E(D \mid \mathrm{do}(X = x'))$, also called the *average causal effect* (ACE). When D is binary, the ACE is the causal risk difference.

We also consider *conditional* causal effects, which for some pre-exposure covariates C are functions of $P(D \mid C; \mathrm{do}(X))$. These denote the causal effect that X has on D within a subgroup defined by $C = c$, which may be different from that in a different subgroup $C = c' \neq c$. Similarly a conditional causal null hypothesis (within $C = c$) is given as

$$H_0^{C=c} : P(D = 1 \mid C = c; \mathrm{do}(X = x)) = P(D = 1 \mid C = c; \mathrm{do}(X = x')) \quad \text{for all } x \neq x',$$

and we define the overall absence of conditional causal effects as

$$H_0^C = \bigcap_{c \in \mathcal{C}} H_0^{C=c}.$$

Remarks:

(1) We require C to be pre-exposure so that $P(D \mid C; \mathrm{do}(X))$ corresponds to a quantity that could in principle be estimated from an RCT, where people with $C = c$ are randomised to different values of X; if C only occurs, or is only defined, after randomisation then this would not be possible. Formally, C being pre-exposure means that these covariates cannot be affected by an intervention, i.e., $P(C \mid \mathrm{do}(X)) = P(C)$. Due to the retrospective nature of case-control sampling, measurements of covariates may (possibly unintentionally) be post-exposure, and in fact measurement of exposure may be post-outcome and hence affected by *recall bias*. Such situations may not only affect the interpretation of the causal parameter but also the validity of causal conclusions due to, e.g., selection bias as explained below.

(2) The causal effects we consider, either as functions of $P(D \mid \mathrm{do}(X = x))$ or $P(D \mid C; \mathrm{do}(X = x))$, are *total* effects in the sense that they measure the causal effect of X on D as possibly 'transmitted' via several pathways. For example, if X indicates smoking then this has a direct negative effect on lung function and therefore health; however taking up smoking may also reduce appetite, and therefore affect health indirectly via body mass index. Breaking an effect down into direct and different indirect causal effects requires the conceptual blocking of some pathways by interventions; this is a significant topic of its own and will not be considered any further here (VanderWeele, 2015).

(3) The above null hypotheses obviously depend on the choice of C. In particular it does *not* hold that $H_0^{\emptyset} \Rightarrow H_0^C$: there could be nonzero effects in subgroups that cancel each other out, so that there is a zero overall effect; this is referred to as *unfaithfulness* (Spirtes *et al.*, 2000). However, for pre-exposure covariates C it does hold that $H_0^C \Rightarrow H_0^{\emptyset}$, so we can regard H_0^C as more informative than H_0.

Even if we are not only interested in testing a causal null hypothesis, it is worthwhile verifying whether a method of estimation is consistent at the causal null under weaker assumptions than general consistency. Indeed, many methods are still valid at the null, even when misspecified models are used; this is also known as the *null-preservation* property (see Chapter 9 of Hernán and Robins, 2018).

(4) Other notions of causal null hypotheses can be found in the literature. For example, when D is continuous one might simply be interested in equality of the means under two treatments $E(D \mid \mathrm{do}(X = 1)) = E(D \mid \mathrm{do}(X = 0))$. A stronger notion is the *sharp* causal null hypothesis which can only be formulated with potential outcomes as it states that there is no causal effect at an individual level, i.e., $D_1^i - D_0^i = 0$ for all units i. We do not consider this further in the present chapter (Hernán, 2004).

6.2.3 Exchangeability and confounding

In order to explain the problem of confounding, we start with the prospective case. Consider a binary treatment X: randomisation of X creates two a priori *exchangeable* groups; that is, at the time of randomisation, those units with $X = 0$ are entirely comparable to units with $X = 1$ regarding the distribution of any measured or unmeasured factors such as age, lifestyle, socioeconomic background. This exchangeability allows causal inference because any subsequent differences between the groups must be due to the different values of X (Hernán and Robins, 2018, Section 3.2). In observational studies exchangeability cannot be guaranteed, as units observed to have $X = 0$ may well have very different attributes than those observed to have $X = 1$; for example, they might be older, have lower alcohol consumption, or higher income. Intuitively, *confounding* occurs when some of these attributes also affect disease risk; formally this implies $P(D \mid X) \neq P(D \mid \mathrm{do}(X))$ as confounding creates a noncausal association between D and X. Hence we seek to take suitable covariates into account to compensate for the lack of exchangeability. A set C of pre-exposure covariates is *sufficient to adjust for confounding* with respect to the effect of X on D (Dawid, 2002; VanderWeele and Shpitser, 2013) if

$$P(D = d \mid C = c; \mathrm{do}(X = x)) = P(D = d \mid C = c, X = x), \tag{6.3}$$

where $P(D \mid C, X)$ is the observational conditional distribution. The absence of confounding corresponds to $C = \emptyset$. In words (6.3) means that once we condition on C, observing $X = x$ is as if X had been set to x by an intervention, at least as far as the resulting distribution of D is concerned; this is called *conditional exchangeability*. Note that, unless we can compare data from an RCT with an observational study, conditional exchangeability is untestable and needs to be justified with subject matter background knowledge. Using potential outcomes, property (6.3) is expressed as $D_x \perp\!\!\!\perp X \mid C$, and also called *strong ignorability* (Rosenbaum and Rubin, 1983).

As explained in the Appendix, with causal DAGs, property (6.3) can be verified by checking that C *blocks all back-door paths* from X to D (Pearl, 1995). In Figure 6.2(a) the only back-door path is given by $X \leftarrow Z \leftarrow C \rightarrow D$ and this is blocked by any of $\{Z\}$, $\{C\}$, or $\{Z, C\}$, but not by $\emptyset$. Minimal sets satisfying (6.3) can be defined, but as the example shows they are not unique (VanderWeele and Shpitser, 2013). Note that if C satisfies (6.3), it does not necessarily follow that a larger set $B \supset C$ satisfies (6.3), even if B consists only of pre-exposure covariates (Greenland, 2003; VanderWeele and Shpitser, 2013).

Methods of adjusting for confounding typically assume that *positivity* is satisfied:

$$0 < P(X = x \mid C = c) < 1 \quad \text{for all } x, c. \tag{6.4}$$

This means that all exposure values are possible for all values of C. As a counterexample,

(a) (b)

FIGURE 6.2
(a) A DAG in which either Z or C alone is sufficient to adjust for confounding. (b) A DAG in which the causal null hypothesis of no effect of X on D holds; including possible ascertainment bias with Z as a symptom.

consider an exposure $X = x_0$ corresponding to two hours of exercise per day; then this is practically impossible to occur for the subgroup of subjects $C = c_0$ who are bedridden. Hence it is not possible to nonparametrically estimate the effect of an intervention $\text{do}(X = x_0)$ for this subgroup as positivity (6.4) is violated.

When C is sufficient to adjust for confounding we have that

$$H_0^C \quad \Leftrightarrow \quad D \perp\!\!\!\perp X \,|\, C, \tag{6.5}$$

which is immediate from the definition (6.3). This implies that any method for testing $D \perp\!\!\!\perp X \,|\, C$ can be used to test H_0^C. Furthermore we obtain the intervention distribution as

$$P(D = d \,|\, \text{do}(X = x)) = \sum_c P(D = d \,|\, C = c, X = x) P(C = c), \tag{6.6}$$

which is known as *standardisation* (Keiding and Clayton, 2014). This means that when C is sufficient to adjust for confounding, $P(D \,|\, \text{do}(X))$ is *identified* by (6.6) if we can obtain data from $P(D \,|\, C, X)$ and $P(C)$, as would be the case in a prospective study with C observed.

However, retrospective data in a case-control study is drawn from $P(D, C, X \,|\, S = 1)$, and neither $P(D \,|\, C, X)$ nor $P(C)$ can necessarily be obtained without additional assumptions. We return to this problem in Sections 6.4 and 6.5.

6.2.4 Selection bias

While confounding is the most well known source of bias affecting causal inference from observational data, a different problem is highly relevant in case-control studies: selection bias (Hernán *et al.*, 2004). As a simple toy example, consider two marginally independent variables $A \perp\!\!\!\perp B$ and assume that they both separately affect a third variable Z such that the joint distribution is given by $P(A, B, Z) = P(A)P(B)P(Z \,|\, A, B)$; graphically we express this as $A \to Z \leftarrow B$. It is easily seen that $P(A, B \,|\, Z)$ does not factorise, so that $A \not\perp\!\!\!\perp B \,|\, Z$ in general. Now, assume that Z indicates being sampled; then we find that two quantities that are not associated in the population may very well be associated in the sample if they both affect the probability of being sampled. For example, if A, B indicate the presence of two diseases that are independent in the population, and $Z = 1$ only occurs if either $A = 1$ or $B = 1$, then A and B will be negatively associated in the subset with $Z = 1$. This kind of problem for medical research has been highlighted by Berkson (1946) in the context of sampling cases and controls from one hospital; he demonstrated that people

FIGURE 6.3
Causal graphs representing forms of recall bias. In (a), some measured covariates C^* are actually recalled after treatment; in (b), we only obtain a measure of exposure after observing the outcome.

who had at least one reason to be in a hospital in the first place are not representative of the population and hence may exhibit 'spurious' associations.

A related example based on Robins (2001) is given in Figure 6.2(b). The DAG shows the null hypothesis of no causal effect of an exposure X (e.g., hormone treatment) on D (e.g., endometrial cancer), and for simplicity there is no confounding affecting X and D. However, exposure may have consequences Z (e.g., bleeding) and subjects with this symptom are more likely to see a doctor. The DAG also expresses that bleeding and cancer may have an unobserved common cause U (e.g., uterine abnormality). In a case-control study, sampling does not only depend on disease status D but also on the presence of symptoms as those are more likely to be found out as cases – this is known as *ascertainment* bias. To confirm this source of possible bias we can apply d-separation to the graph (see Appendix), and find that even though there is a marginal independence $D \perp\!\!\!\perp X$ in the whole population, there is no such independence in the sampled population: $D \not\perp\!\!\!\perp X \mid S$; in addition, this cannot be corrected by stratifying on Z because $D \not\perp\!\!\!\perp X \mid (Z, S)$. The reason for the selection bias here is that when Z affects the probability of being sampled, then X and U become dependent in the sample (as in the earlier toy example we could consider the extreme case where $Z = 1$ if either $X = 1$ or $U = 1$) which creates a noncausal association between X and D. This example illustrates how selection bias can impede identification of the causal null hypothesis (see Section 6.3.1); even if there is a nonzero causal effect of X on D, using DAGs in this way alerts us to the possibility that its estimation might be biased.

6.2.5 Post-exposure covariates and recall bias

An issue closely related to selection bias and the retrospective nature of the sampling is that the actual measurements of covariates and exposure may not follow the intended time ordering (remember BH's criterion of 'temporality'). For example, we may want to adjust for baseline BMI, but are only able to obtain a BMI measurement from after exposure, perhaps because it was taken after first symptoms occurred; alternatively, a subject may try to recall their BMI from before exposure, but may not do so accurately. Similarly, exposure itself may be difficult to retrieve from an objective source, and is instead measured by asking the patient; again, the *recalled* value of exposure may be systematically different for cases and controls.

Let us consider the post-exposure covariate problem first; a possible scenario is depicted in Figure 6.3(a). While the genuine pre-exposure covariate C would be sufficient to adjust for confounding, a post-treatment version of it (C^*) is not. This is due to various issues reflected in certain paths in the DAG: first, the path $X \leftarrow C \rightarrow D$ means there is residual confounding by C, for which C^* cannot adjust; second, the path $X \rightarrow C^* \rightarrow D$ implies that

part of the causal effect is transmitted via C^*, so that we should not adjust for it if we desire a total effect for inference; third, the path $C \to C^* \leftarrow U \to D$ means that conditioning on C^* creates a spurious association between the unobserved U and the original C (as in our toy example above) resulting in a new source of residual confounding. Note that one might think that a different reason for conditioning on C^* is the desire to estimate a 'direct' effect of X on D, but again the presence of an unobserved common cause U leads to a noncausal (X, D)-association and hence bias from a selection effect due to $X \to C^* \leftarrow U \to D$; see Aschard *et al.* (2015) for an illustration of this bias in the context of GWAS case-control studies.

Figure 6.3(b) shows the case of the exposure X being affected by recall bias, because the *measured* exposure X^* depends on D; this can occur if $D = 1$ is a negative health outcome and leads patients to search for explanations, and hence exaggerate the actual exposure in their recollection. Nonparametrically, it is impossible to draw causal conclusions based on only D, X^*. If the actual exposure X can be retrieved, we see from the DAG that X^* should be ignored; adjusting for it would lead to a noncausal association because of the selection effect. More generally, there could be other reasons leading to the measured exposure X^* being partly affected by D, suggesting an association which is really due to reverse causation. This can happen, e.g., if both exposure and outcome reflect underlying processes that have developed over time such as cardiovascular problems leading to increased fibrinogen levels instead of the other way around.

6.2.6 Nonparametric identifiability

A case-control study provides data from the observational (i.e., noninterventional) distribution $P(X, C, D \mid S = 1)$. The key problem of *causal inference* is: under what assumptions (if at all) are certain aspects of the intervention distributions $P(D \mid \mathrm{do}(X))$ or $P(D \mid C; \mathrm{do}(X))$ uniquely computable, i.e., *identified*, from $P(X, C, D \mid S = 1)$. For instance in the example of ascertainment bias given earlier, the marginal independence corresponding to a causal null hypothesis cannot be identified from the conditional distribution. We mainly consider *nonparametric identifiability*, relying only on structural assumptions that can be expressed in DAGs (Pearl, 2009), and not particular parametric choices (but see Section 6.5). Moreover, we mainly consider the case of no external information such as the disease prevalence $\tilde{p} = P(D = 1)$, which typically means that besides the causal null hypothesis, only the conditional causal odds ratio (COR) is identified as addressed below.

6.3 Identifying the causal null hypothesis and odds ratios

Let us start by considering the simplest situation. Assume sampling depends on disease status alone (6.1) and there is no confounding, i.e., $P(D \mid \mathrm{do}(X = x)) = P(D \mid X = x)$. Hence, we can identify any causal effect measure if we can obtain $P(D \mid X)$ from $P(X, D \mid S = 1) = P(X \mid D)P(D \mid S = 1)$. However, $P(D \mid X)$ cannot be obtained from $P(X \mid D)$ without additional information such as the disease prevalence $\tilde{p}$ in the population, so that the causal risk difference or risk ratio are not identified. Instead, we can assess from $P(X \mid D)$ whether X and D are independent, and we can compute the (X, D)-OR. Therefore, we start by focusing on approaches to tackle two specific types of causal identification questions: Is there a causal effect of an exposure X on a disease D at all? In other words, can we test a causal null hypothesis? What is the COR describing the causal effect of X on D? It turns out that these questions are closely related and rely on the same assumptions.

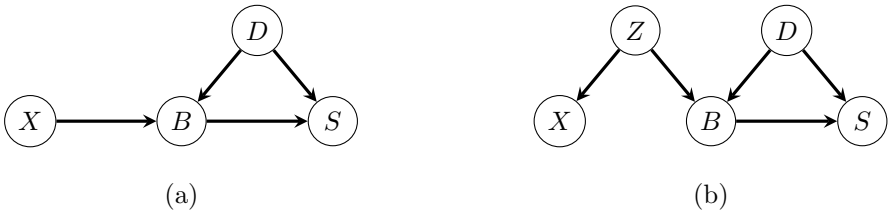

FIGURE 6.4
Examples of possible selection bias: (a) presence in database B directly depends on exposure; (b) case where covariates Z can be found to explain presence in database.

6.3.1 Selection bias and the causal null hypothesis

As alluded to above, if we merely want to know whether X and D are associated at all, it does not matter if we observe data from $P(X \mid D)$ rather than $P(D \mid X)$, because $D \perp\!\!\!\perp X$ can be established by checking that $P(X \mid D) = P(X)$. The problem of testing a causal null hypothesis is similar, except for the need to consider confounding. Due to property (6.5), when C is sufficient to adjust for confounding $H_0^C \Leftrightarrow D \perp\!\!\!\perp X \mid C$, and this independence can be checked using either $P(X, C \mid D)$ or $P(X \mid C, D)$. A case-control study provides data from $P(X, C, D \mid S = 1)$; hence it is sufficient to assume (6.2), so that $P(X, C, D \mid S = 1) = P(X \mid C, D)P(C, D \mid S = 1)$. In fact, assumption (6.2) means that selection bias is ruled out, because it says that the sampling does not, directly or indirectly, depend on both the exposure and the disease status, possibly after taking covariates into account.

As seen in the ascertainment bias example, a graphical 'trick' to check for selection bias is to draw a DAG showing the structural assumptions under the null hypothesis (Hernán *et al.*, 2004; Didelez *et al.*, 2010b; Richardson and Robins, 2013); the latter typically means that there is no directed edge nor directed path from X to D. We can then check graphically whether a set C exists such that each remaining path between X and D is blocked by (C, S); here S needs to be included because the design implies that everything is conditioned on S. If no such set of observable variables C can be found (e.g., when only C^* and not C is available in Figure 6.3(a)), then it is typically impossible (without further information or assumptions) to test the causal null hypothesis. In other situations it may be possible to gather additional information so as to enable valid inference. Consider for example the DAG in Figure 6.4(a). The absence of a directed edge from X to D represents the causal null hypothesis, and for simplicity we assume no confounding. Further, the DAG represents the situation where controls are found in some database (with binary variable B indicating presence in the database), but where exposure leads to under- or over-representation in that database. If these are the only observable variables, then the causal null hypothesis cannot be tested from data conditional on $S = 1$, i.e., it is not identified. However, assume instead the DAG in Figure 6.4(b); this may have been obtained after investigating exactly how subjects get into the database and it is found that a set of covariates Z both explains the presence in the database and predicts exposure X such that $X \perp\!\!\!\perp B \mid Z$. We now find that $S \perp\!\!\!\perp X \mid (Z, D)$. While Z is in this case not needed to adjust for confounding, it does not violate (6.3). Hence, we can test the causal null hypothesis H_0^Z from the sampled data.

For example, suppose B is a register of workers at a nuclear power plant, X is a measure of radiation exposure, and D is an indicator of cancer. Workers may be added to the register either as a precaution based on their level of radiation exposure, or because they have been diagnosed with cancer. This means that selecting individuals from the register for a study may introduce a spurious negative association between X and D, because they

are competing explanations for inclusion in the register. However, if precautionary inclusion in the register was based on the individual's role at the plant, Z, as in Figure 6.4(b), then using Z could allow recovery from selection bias.

The question of whether a conditional independence can be tested from a selected sample turns out to be closely related to the question whether the corresponding OR is collapsible. This is therefore addressed next.

6.3.2 Collapsibility of odds ratios

The odds ratio is, first and foremost, a measure of association; in the following section we explore under what assumptions and in what sense it can be used for causal inference. Here we revisit some of its basic properties that will be needed later. We state results below for binary X and D; for generalisation to the nonbinary case see Didelez *et al.* (2010b).

For a binary outcome D and binary exposure X we define the odds ratio OR_{DX} as

$$OR_{DX} = \frac{P(D=1\,|\,X=1)P(D=0\,|\,X=0)}{P(D=0\,|\,X=1)P(D=1\,|\,X=0)}.$$

The OR has a number of (related) properties that render it very useful in case-control studies: (i) it is symmetric: $OR_{DX} = OR_{XD}$; (ii) it is invariant to changes in the marginal distributions of D and X; (iii) it can be used to test independence, since $D \perp\!\!\!\perp X \Leftrightarrow OR_{DX} = 1$. With (i) and (ii) it follows in particular that the OR can be identified based on data from the conditional distribution $P(X\,|\,D)$ even when the sampling proportion of cases $\tilde{p}_S$ is not equal to the population proportion of cases $\tilde{p}$.

We extend the definition to take (a set of) covariates C into account: the conditional odds ratio $OR_{DX}(C=c)$ is given by

$$OR_{DX}(C=c) = \frac{P(D=1\,|\,X=1,C=c)P(D=0\,|\,X=0,C=c)}{P(D=0\,|\,X=1,C=c)P(D=1\,|\,X=0,C=c)}.$$

We write $OR_{DX}(C)$ to refer to the set of all ORs $\{OR_{DX}(C=c), c \in \mathcal{C}\}$. The conditional OR is also symmetric; is invariant to the marginal distributions of $X\,|\,C$ and $D\,|\,C$; and $OR_{DX}(C=c) = 1 \Leftrightarrow D \perp\!\!\!\perp X\,|\,C=c$.

To define collapsibility, assume that C consists of at least two variables C_1, C_2 (each can be a vector). Then $OR_{DX}(C_1, C_2)$ is *collapsible* *over* C_2 if for all $c_2 \neq c_2'$ and all c_1 in the respective domains

$$OR_{DX}(C_1=c_1, C_2=c_2) = OR_{DX}(C_1=c_1, C_2=c_2') = OR_{DX}(C_1=c_1). \tag{6.7}$$

A sufficient condition for property (6.7) is that either $D \perp\!\!\!\perp C_2\,|\,(C_1, X)$ or $X \perp\!\!\!\perp C_2\,|\,(C_1, D)$, and this becomes necessary when C_2 is binary (Whittemore, 1978).

It is interesting to note that in an RCT where X is randomly assigned and hence marginally independent of any covariates C, it is not generally true that $OR_{DX}(C)$ is collapsible over C even when the first equality of (6.7) holds for all of C (i.e., there is no effect modification by C); this is because the marginal independence $X \perp\!\!\!\perp C$ does not help with collapsibility. This may be regarded as a disadvantage of using the OR as a measure of association or effect measure; however, in such a case collapsing over C simply leads to an attenuation towards one (Zeger *et al.*, 1988).

6.3.3 Identification – main results

We define the (conditional) *causal* odds ratio, $COR_{D|X}(C = c)$, for pre-exposure covariates C as

$$COR_{D|X}(C = c) = \frac{P(D = 1 \,|\, C = c; \mathrm{do}(X = 1))P(D = 0 \,|\, C = c; \mathrm{do}(X = 0))}{P(D = 0 \,|\, C = c; \mathrm{do}(X = 1))P(D = 1 \,|\, C = c; \mathrm{do}(X = 0))}.$$

Remarks:

(1) The COR is not symmetric in (X, D), as causality, unlike stochastic dependence, is not symmetric; if X is causal for D then D is not causal for X, i.e., if $COR_{D|X}(C = c) \neq 1$ then $COR_{X|D}(C = c) = 1$.

(2) As noted earlier, even if $COR_{D|X}(C = c) = COR_{D|X}(C = c')$, for all $c \neq c'$, we do not generally have $COR_{D|X}(C) = COR_{D|X}$; the latter is attenuated towards the null. However, $COR_{D|X}(C) \equiv 1$ implies $COR_{D|X} = 1$.

(3) The CORs are straightforward to generalise to nonbinary X, in which case it is common to choose a reference category and compute a set of CORs.

The main implication of Sections 6.3.1 and 6.3.2 is that to test the causal null hypothesis H_0^C and estimate the CORs we need to be able to collapse over S. Throughout we assume that $S \not\perp\!\!\!\perp D \,|\, (C, X)$, i.e., that the sampling is dependent on the outcome variable D as would be the case in any case-control study. Then we have the following result (more general results and proofs are given in Didelez *et al.* (2010b))

Proposition 1 *(i) The conditional odds ratio $OR_{DX}(C, S)$ is collapsible over S if and only if (6.2) holds; hence $OR_{DX}(C, S) = OR_{DX}(C)$. Further, if C is sufficient to adjust for confounding w.r.t. the effect of X on D then the conditional OR is equal to the causal OR, $OR_{DX}(C) = COR_{D|X}(C)$.*
(ii) If (6.2) holds then

$$D \perp\!\!\!\perp X \,|\, C \Leftrightarrow D \perp\!\!\!\perp X \,|\, (C, S = 1),$$

which allows us to test the left-hand side conditional independence even under selection, i.e., conditional on $S = 1$. Further, if C is also sufficient to adjust for confounding of the effect of X on D then testing for independence under selection, $D \perp\!\!\!\perp X \,|\, (C, S = 1)$, is a test of the causal null hypothesis H_0^C.

Remarks:

(1) In Proposition 1, part (ii) is not restricted to the situation of categorical variables so that whatever test statistic seems appropriate given the measurement scales of X, D, C can be used.

(2) In the particular case of binary D and continuous X it is well known that we can consistently estimate the OR using a logistic regression (Prentice and Pyke, 1979). This result, however, relies on the logistic link being correctly specified, while the above results for categorical X and D make no parametric assumptions.

(3) While for a causal interpretation C needs to be sufficient to adjust for confounding, it also needs to contain any matching variables as otherwise (6.2) is unlikely to be satisfied.

(4) Proposition 1 can be generalised in that it is possible that variables that are themselves affected by exposure can sometimes be used to obtain causal conclusions. The key here is that if X affects Z which in turn affects S, it can sometimes be possible to first collapse over S given Z, and subsequently collapse over Z to obtain a COR. The required conditional independence assumptions can again easily be checked graphically on DAGs. An example in the context of retrospective, but not case-control, sampling is given in Didelez *et al.* (2010b); Bareinboim and Pearl (2012) give a complete identification algorithm.

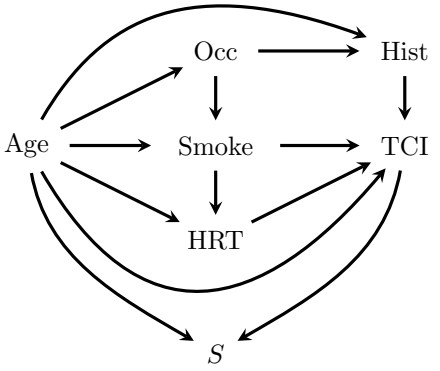

FIGURE 6.5
Example DAG for case-control study on the effect of HRT on TCI; matching is on Age; other covariates: Occupation, Thromboembolic history, Smoking.

Example: Causal effect of HRT on TCI

The above is illustrated with a study of the effect of hormone replacement therapy (HRT) on transient cerebral ischemia (TCI) (Didelez *et al.*, 2010b); note that this is a simplified version of the actual analysis by Pedersen *et al.* (1997). The study is a case-control study matched by age, so that $S \perp\!\!\!\perp HRT \,|\, (Age, TCI)$ but not $S \perp\!\!\!\perp HRT \,|\, TCI$; hence the marginal HRT-TCI OR cannot be consistently estimated as it is not collapsible over S, but the OR conditional on 'age' can consistently be estimated from the sampled data. However, this conditional OR does not necessarily have a causal interpretation unless 'age' is sufficient to adjust for confounding. To examine this, further covariates 'smoking', 'history of thromboembolic disorders' (TH), and 'occupation' were included in a DAG to represent the structural assumptions (Figure 6.5). It can be seen that covariates 'smoking' and 'age' are sufficient to adjust for confounding. Hence the conditional HRT-TCI ORs given these two covariates do have a causal interpretation, and if significantly different from one allow us to reject the causal null hypothesis. For instance the OR of taking oestrogens compared to no HRT was estimated to be just above 2 (pooled over categories of 'smoking' and 'age') which suggests that there is a causal effect of taking oestrogens resulting in higher risk of TCI.

6.4 Reconstructing the joint distribution

In this section, we address the question of when the joint population distribution $P(D, X, C)$ or possibly $P(D \,|\, X, C)$ is identified from a case-control design, i.e., from $P(D, X, C \,|\, S = 1)$. This is especially relevant if we wish to estimate causal effect measures other than conditional ORs, such as unconditional causal risk differences or risk ratios which are more suitable for comparison with results from RCTs. The results of Bareinboim and Tian (2015) and Bareinboim *et al.* (2014) suggest a pessimistic answer to this, as they demonstrate that the conditional distribution $P(Y \,|\, T)$ can be recovered from $P(Y \,|\, T, S = 1)$ if and only if $Y \perp\!\!\!\perp S \,|\, T$; if we take $Y = D$ then this condition will never be satisfied in a case-control

FIGURE 6.6
Examples where independencies allow recovery of joint distributions from data under selection bias: (a) marginal independence of X and G; (b) conditional independence of X and G given Z.

study, where selection S always depends on D by design. Still, we can make some headway if we somewhat relax the definition of 'identifiability' as demonstrated by Evans and Didelez (2015). This is best explained with a small example.

Example: Case-control study with genetic covariate

Assume we are interested in the causal effect of an exposure X (binary or categorical) on disease D and we have additional measurements on genotype G which is *known* to be independent of X in the population, i.e., $X \perp\!\!\!\perp G$. The graphical representation of this marginal independence is via a so-called 'V-structure' where arrows meet head-to-head (at D) with no edge between the parents, X and G, as illustrated in Figure 6.6(a). Such studies are for example carried out to investigate gene-environment interactions (Moerkerke *et al.*, 2010). Further assume that the case-control design is such that $P(X, G, D \mid S = 1) = P(X, G \mid D)P(D \mid S = 1)$. As $X \perp\!\!\!\perp G$, we know that the joint distribution factorises $P(X, G) = P(X)P(G)$; hence we must have

$$\sum_{d=0,1} P(X = x, G = g \mid D = d)P(D = d)$$
$$= \sum_{d=0,1} P(X = x \mid D = d)P(D = d) \sum_{d=0,1} P(G = g \mid D = d)P(D = d)$$

for each value of x, g. This results in (at most) $(|\mathcal{X}| - 1)(|\mathcal{G}| - 1)$ nonredundant equations, each of which is quadratic in the unknown $\tilde{p} = P(D = 1)$ and can have at most two solutions. The true (population) distribution of D must be a solution to each equation, so it is identified up to at most two solutions. We note, however, that the equations become uninformative (i.e., any distribution for D is a solution) if $G \perp\!\!\!\perp D \mid X$ or $X \perp\!\!\!\perp D \mid G$.

General result

The above example demonstrates that if certain marginal independencies (in the population) are known, then we may be able to reconstruct the joint distribution from a conditional one; however, there may be more than one solution, and even infinitely many solutions if there are 'too many' independencies in the population. We call this *generic* identifiability (Evans and Didelez, 2015). Formally, generic k-identifiability means that there are at most k solutions under any true distribution in the model class except possibly for a proper (i.e., lower dimensional) algebraic subset of the model class. The existence of solutions depends on the number of known constraints being at least as large as the number of unknowns,

which in an unmatched case-control study would be just one: $\tilde{p}$. The number of constraints in the above example is $(|\mathcal{X}|-1)(|\mathcal{G}|-1)$, so if both X and G are binary, there is exactly one constraint.

Entirely general results are difficult to obtain in this context. Here we provide a relatively general result modified from Evans and Didelez (2015) to reflect the situation of a case-control study. Hence, D is binary; we assume that selection is based only on D, so (6.1) and therefore data is available from $P(X, C \mid D)$. The covariates C can be a vector of variables which would correspond to a set of nodes in a DAG.

Proposition 2 *Assume $P(X, C \mid D = d)$, $d = 0, 1$ is given, and the joint population distribution $P(X, C, D)$ obeys a known DAG.*
(i) A necessary (but not sufficient) condition for $P(X, C, D)$ to be generically identifiable is that the DAG contains at least one V-structure.
(ii) If the graphical parents of D are given by $\{X, C\}$, then $P(X, C, D)$ is generically identifiable if at least one constraint is implied on $P(X, C)$ by the DAG.

To illustrate the above result consider the example in Figure 6.6(b) with $C = G \cup Z$. As in Figure 6.6(a) we have a V-structure $X \to D \leftarrow G$, so property (i) of Proposition 2 is satisfied. The parents of D are G, Z, X, and from the DAG we can read off that $X \perp\!\!\!\perp G \mid Z$ which induces at least one constraint, so property (ii) of Proposition 2 is satisfied and generic identification holds.

Remarks:

(1) Proposition 2 allows causal inference if the causal effect of interest is further identified from $P(X, C, D)$. So for instance when C is sufficient to adjust for confounding we can use the identity (6.6) to compute $P(D \mid \mathrm{do}(X = x))$.

(2) Before quantifying a causal effect, one may first want to test the null hypothesis of no (conditional) causal effect, H_0^C. As noticed before, when the null translates into a conditional independence (or absence of a directed edge in a DAG), this may actually hamper identification. We conjecture that in these cases the null can always be checked directly from a factorisation of $P(X, C \mid D)$ itself; Proposition 1 is an example of this.

(3) When the number of constraints in $P(X, C)$ is larger than one, the model (DAG) itself becomes empirically testable. For instance in Figure 6.6(b) with all variables binary, there are two constraints: $X \perp\!\!\!\perp G \mid Z = z$ for $z = 0, 1$. This yields two equations for one unknown, $\tilde{p}$; so if the two solutions do not agree the model assumptions must be violated.

(4) Once generic identification is established, it is in principle straightforward to fit the model. To perform maximum likelihood estimation we need to maximise the conditional log-likelihood based on $P(X, C \mid D)$, e.g., with the TM algorithm of Edwards and Lauritzen (2001).

(5) Practical problems may occur: if certain associations between (some elements of) C, X, and D are weak, identification becomes unstable. Additionally, if D represents a very rare disease, then the marginal distribution of (X, C) is approximately the same as that for the controls, $P(X, C) \approx P(X, C \mid D = 0)$, so recovering $P(X, C)$ exactly may not lead to substantial new insights. However, the method could be used for sensitivity analyses in such situations.

6.5 Adjusting for confounding

The previous sections addressed nonparametric approaches to causal inference, in that we used only structural assumptions and no external information to obtain identification. These run into practical problems if there are many covariates or some variables are continuous. Hence, we now discuss how popular (semi-)parametric methods for adjusting for confounding with prospective data can be adapted to case-control studies.

For a causal interpretation of the results, we see from the previous sections that we need two assumptions, which we use throughout:

(i) the covariates C are sufficient to adjust for confounding; and

(ii) sampling satisfies (6.2)

so that there is no selection bias and we can collapse over S; note that this typically means that C must contain any covariates used for matching.

6.5.1 Regression adjustment

Regression adjustment simply includes the confounders C into the outcome regression, which means positing a model for $P(D|X, C)$. Note that this results in conditional causal effect measures, which only equal the marginal ones if they are collapsible.

For case-control data, standard theory suggests that, even under retrospective sampling, a correctly specified logistic regression of D on (X, C) consistently estimates the conditional (X, D)-ORs, with only the intercept of that regression distorted due to the sampling proportions being different from the population ones (Prentice and Pyke, 1979). Then, under the assumptions above, the logistic regression coefficients correspond to *conditional* causal (log) ORs. When all variables are categorical and a saturated logistic regression is used, this approach is in fact identical to the one in Section 6.3.3. However, if some of (X, C) are continuous or if higher-order interaction terms are omitted from the logistic regression, this model imposes parametric assumptions on which the validity of the inference rests.

If the logistic regression is misspecified, the approach typically results in biased effect estimates. Unfortunately, this can also happen under the null hypothesis H_0^C, i.e., null-preservation is not guaranteed meaning that the coefficient(s) for the exposure might not be zero under H_0^C if the model is misspecified. One could therefore supplement the analysis by other tests of this null hypothesis along the lines of Proposition 1.

6.5.2 Standardisation

Standardisation is based on equality (6.6), where typically the terms required for the right-hand side are estimated and then plugged in to obtain the left-hand side. This again relies on a parametric model for $P(D \mid X, C)$. However, unlike regression adjustment, standardisation yields the marginal intervention distribution $P(D \mid \mathrm{do}(X))$, from which any marginal causal effect parameters can be computed.

To obtain all ingredients for the right-hand side of (6.6) from case-control data requires additional external information. A number of relevant methods in this context using additional knowledge on the population prevalence $\tilde{p} = P(D = 1)$ are discussed in Persson and Waernbaum (2013). With $\tilde{p}$ we can obtain exact weights to adjust the intercept of the logistic regression model yielding correct predicted values for $P(D = d \mid C = c, X = x)$; further $\tilde{p}$ helps to obtain the case-control weighted distribution of the covariates corresponding to $P(C = c)$. This results in the following estimate of the intervention distribution obtained

as weighted average over the covariates

$$\begin{aligned}\hat{P}(D = 1 \mid \mathrm{do}(X = x)) \\ = \quad \frac{\tilde{p}}{\#\text{cases}} \sum_{\text{cases}} \operatorname{expit} g(x, C_i; \hat{\beta}^*) + \frac{1-\tilde{p}}{\#\text{controls}} \sum_{\text{controls}} \operatorname{expit} g(x, C_i; \hat{\beta}^*)\end{aligned}$$

where $g(\cdot)$ is the chosen linear predictor and $\hat{\beta}^*$ is the vector of estimated logistic regression coefficients including the adjusted intercept. The resulting $\hat{P}(D = 1 \mid \mathrm{do}(X = x))$ can then be summarised as marginal causal parameter, such as a marginal COR or risk ratio.

An alternative to adjusting the intercept is to weight the likelihood equations according to the above weights for cases and controls and obtain $\hat{\beta}_w$ as estimated coefficients to be inserted in the above formula instead of $\hat{\beta}^*$. For matched case-control data all of the above estimators can still be applied, but with modified weights that require the prevalence within matching values, i.e., $P(D = 1 \mid M = m)$, where $M \subset C$ are the matching variables (Persson and Waernbaum, 2013). In case that prevalence information is not available (for the un-matched or matched estimators), the authors recommend a sensitivity analysis. Their comparative simulation study suggests that all estimators perform well at finite sample sizes when the logistic regression is correctly specified, but exhibit considerable bias if a misspecified model, e.g., a probit instead of a logit link, is used; sensitivity to misspecified weights is not explored though. Evidently, the same potential lack of null preservation holds as mentioned earlier if the logistic regression is misspecified.

A different approach still relying on $\tilde{p}$ is Targeted Maximum Likelihood estimation (TMLE) (van der Laan, 2008); this additionally exploits the propensity score which we address further below. The TMLE remains consistent if the propensity score is correctly specified even if the logistic regression is not, i.e., it is *doubly robust*.

6.5.3 Propensity scores

For binary exposure, the propensity score is given by $\pi = P(X = 1 \mid C)$; methods for nonbinary exposure exist, but we refrain from going into these here in order to focus on the basic principle. Note that due to assuming positivity (6.4) we have that $0 < \pi < 1$. The propensity score has the property that $X \perp\!\!\!\perp C \mid \pi$, known as *balancing* of the covariates between treatment groups (Rosenbaum and Rubin, 1983). As a consequence, if C is sufficient to adjust for confounding then so is π. The underlying principle can be illustrated with the DAG in Figure 6.2(a), letting $Z = \pi$.

Two related ways of using the propensity score to adjust for confounding with prospective data are by matching or stratification on π, and regression adjustment using the propensity score as covariate in addition to X. Stratification and regression adjustment result in conditional causal effect measures (conditional on π).

A third way of using π is known as inverse probability of treatment weighting (IPTW) (Robins *et al.*, 2000). With prospective data, the method proceeds by re-weighting the sampled units by $\pi_i = P(X = 1 \mid C = c_i)$ if $X_i = 1$ and by $1 - \pi_i = P(X = 0 \mid C = c_i)$ if $X_i = 0$, and then fit a parametric or semi-parametric model for $P(D \mid X)$ on the weighted data. It can be shown that X and C are independent in the weighted sample; IPTW can therefore be thought of as 'removing' any incoming edges from C into X in a causal DAG. Hence, if all models are correctly specified (and under (6.3)), IPTW consistently estimates a *marginal structural model (MSM)*, i.e., a model for $P(D|\,\mathrm{do}(X))$ (Robins *et al.*, 2000). So, like standardisation, IPTW yields marginal causal effect parameters.

In observational studies π is often estimated, typically by fitting a model to $P(X = 1 \mid C)$. However, in a case control setting where everything is conditional on $S = 1$, we do not typically have direct access to data from $P(X = 1 \mid C)$. Månsson *et al.* (2007) provide an

overview and comparison of possible approaches for obtaining an estimate of π. In some case-cohort studies, a sub-cohort may be available from which π can consistently be estimated, or if the sampling fraction is known it can exactly be re-weighted; both these approaches result in consistent estimation of a causal effect if the involved models are correctly specified. Other approaches consist of fitting a model for π on the controls only, or on the whole sample; or fitting a model for $P(X = 1 \mid C, D)$ and predicting X for each individual as if they were a control. These last three methods all have the null preserving property that if the model for π is correct they consistently estimate π if X has no effect on D, but they are not consistent otherwise. Estimating π from the controls only is 'nearly' consistent if the disease is very rare. A lack of consistency of the propensity score estimators means that there will be some residual confounding so that we must expect the estimator of the causal effect to be biased. This bias, though small, is confirmed by the simulations of Månsson *et al.* (2007). Further, in finite samples, their simulations suggest that propensity score stratification using the above methods for case-control data sometimes results in an artificial effect modification by the propensity score even under the null and without confounding.

A general issue arising when using the propensity score for stratification or regression adjustment with case-control data is that the outcome model is typically a logistic regression; but as shown in Section 6.3.2 in a DAG such as Figure 6.2(a) with $Z = \pi$, the $OR_{DX}(\pi, C)$ is not collapsible over C as $X \not\perp\!\!\!\perp C \mid (D, \pi)$. Even in the absence of effect modification by C it is possible that $COR_{DX}(\pi)$ is different from $COR_{DX}(C)$. Furthermore, in the presence of effect modification by C, it appears more difficult to interpret $OR_{DX}(\pi)$ meaningfully.

6.6 Instrumental variables

All previous sections relied on the availability of sufficient covariates to control for confounding. Now we consider an approach that can sometimes be employed when this is not the case, i.e., when there is *unobserved* confounding. An *instrumental variable* (IV) can be regarded as 'imperfectly mimicking' randomisation of X when exposure itself cannot be randomised. When the IV is a genetic variant this has become known as Mendelian randomisation (MR) (Davey Smith and Ebrahim, 2003; Didelez and Sheehan, 2007a). Case-control studies play an important role for research into Mendelian randomisation, as the method is often applied in a secondary or meta-analysis, where (some of) the primary analyses were case-control studies. A brief overview of the use of IVs for case-control data is also given in Windmeijer and Didelez (2018).

6.6.1 Definition of instrumental variables

A variable Z is an instrument for the causal effect of an exposure X on an outcome D in the presence of unobserved confounding U under the following three core conditions (Didelez and Sheehan, 2007a).

IV.1: $Z \perp\!\!\!\perp U$, the instrument is independent of the unobserved confounding;
IV.2: $Z \not\perp\!\!\!\perp X$, the instrument and exposure are dependent;
IV.3: $D \perp\!\!\!\perp Z \mid (X, U)$, the outcome is conditionally independent of the instrument.

Here U is such that if it were observed it would be sufficient to adjust for confounding (for simplicity we ignore observed covariates). Note that Z can be vector valued, such as a set of genetic variants obtained from genome wide association studies (GWASs).

FIGURE 6.7
DAGs representing (a) the IV conditions (prospectively); (b) the IV conditions under the causal null, with a sampling indicator.

The IV assumptions are represented graphically in Figure 6.7(a); it is the absence of any edges between Z and U, as well as between Z and D, that, respectively, implies conditions IV.1 and IV.3. In particular, the assumptions imply that there are no common causes of Z and U, nor of Z and D, which means that we can treat Z in the same way we treat a genuinely randomised quantity; moreover, the only 'input' that Z has into the system is via X, so that we can regard Z as a proxy for randomisation of exposure X. Intuitively, the stronger the dependence between Z and X the better, since this reduces X's susceptibility to being affected by unmeasured variables U; this is expressed in condition IV.2.

An IV can help causal inference in the form of testing and bounding causal quantities, and (under additional assumptions) point estimation. In the following we discuss how to adapt the standard approaches to a case-control setting, where we only observe data from $P(Z, X, D \mid S = 1)$. We assume no matching (we still ignore any observed covariates C), and also that sampling only depends on the outcome: $S \perp\!\!\!\perp (X, U, Z) \mid D$. Together with the IV conditions this implies $P(Z, X, D \mid S = 1) = P(Z, X \mid D)P(D \mid S = 1)$.

6.6.2 Testing the causal null hypothesis with an instrumental variable

With prospective data, we can test for the presence of a causal effect of X on D by testing for an association between Z and D. This is because the only way that Z and D can be associated is via a causal effect of X on D; when we delete the edge from X to D in Figure 6.7(a), the only remaining path $Z \to X \leftarrow U \to D$ is unconditionally blocked, so there is no (Z, D)-association. Of course, if the dependence between Z and X (condition IV.2) is only weak relative to the sample size, then this will not be a powerful test.

A small caveat applies to this test if the unobserved confounders U are likely to contain effect modifiers such that the effects in different subgroups have different directions; for example, a treatment which is beneficial for men but harmful for women. These could in principle cancel out exactly, leading to an unfaithful distribution. For many practical situations we consider this to be of minor relevance (Didelez and Sheehan, 2007a).

The null hypothesis (absence of the edge $X \to D$) under case-control selection is illustrated in Figure 6.7(b). We see that case-control sampling, i.e., conditioning on $S = 1$, produces no association between Z and D under the null as the only path between them is blocked by S. Hence, in analogy to the prospective case, we find that testing $Z \perp\!\!\!\perp D \mid S = 1$ with case-control data is a valid test for causal effect of X on D, where strictly speaking we also assume faithfulness.

As an example consider the analysis of Meleady *et al.* (2003). The data comes from a case-control study and includes homocysteine levels X, cardiovascular disease D (CVD), as well as a genetic IV (the rs1801133 polymorphism) Z with three levels. A χ^2-test of the

association between Z and D yields a p-value of 0.357 so no evidence for an association. If Z satisfies the IV conditions this means no evidence for a causal effect of homocysteine on CVD. Note that as this test does not make use of data on the exposure X, it can be carried out using case-control data from, say, a separate study with data on Z and D only.

6.6.3 Instrumental variable estimation

Without additional assumptions, the IV conditions do not generally allow us to identify the causal effect of X on D. However, in the all-discrete case, they impose bounds on the causal effect in the form of minima and maxima for $P(D = 1 \mid \mathrm{do}(X = x)), x = 0, 1$ (Balke and Pearl, 1994). These bounds are given as simple transformations of the observed relative frequencies and can for example be computed with the Stata command `bpbounds`. For examples including an application to the above CVD case-control data, see Palmer *et al.* (2011). When applicable, it is recommended to compute the bounds, as they demonstrate what information the data provide without additional parametric constraints.

Point estimation with an IV requires additional parametric assumptions. The simplest case is given by a linear model which gives rise to the *Two-Stage-Least-Square (TSLS)* estimator (Wooldridge, 2010). Even though this is not a suitable assumption in a case-control situation, we mention it here as it has inspired OR estimation which is relevant. Assume that

$$E(D \mid X, U) = \theta X + h(U).$$

Under IV.1-IV.3 it can be shown that $\theta = \mathrm{Cov}(D, Z)/\mathrm{Cov}(X, Z)$, so a consistent estimator $\hat{\theta}_{IV}$ for θ is given by the ratio of the regression coefficient $\hat{\beta}_{D|Z}$ from a regression of D on Z, and $\hat{\beta}_{X|Z}$ of X on Z; that is

$$\hat{\theta}_{IV} = \frac{\hat{\beta}_{D|Z}}{\hat{\beta}_{X|Z}}, \tag{6.8}$$

where Z, X, D are all univariate. More generally, it can be shown that the above is a special case of TSLS, which consists of regressing X on Z to obtain predicted values $\hat{X}$, and then regressing D on $\hat{X}$. From the above, we see that condition IV.2 needs to be strengthened to $\mathrm{Cov}(X, Z) \neq 0$. In fact, when $\mathrm{Cov}(X, Z)$ is close to zero, the TSLS estimator is not only unstable but biased, a phenomenon known as weak-IV bias (Wooldridge, 2010).

Turning to case-control data, note that when we assumed 'no unobserved confounding' in earlier sections of this chapter; targeting an OR as the parameter of interest had the advantage that it often does not require additional information. However, this advantage is lost with IVs and case-control data. If we want to estimate a COR with an IV we still need external information on the sampling proportions or disease prevalence. In the absence of such information, only approximate solutions are available, which we address first.

6.6.3.1 Instrumental variable estimation without external information

Inspired by the linear case where a simple IV estimator of the causal effect is given by the ratio (6.8), it has been suggested to estimate the causal log-OR for the effect of X on D as the ratio of the log-OR between outcome and IV divided by the linear regression coefficient for X on the IV Z, $\beta_{X|Z}$ (Casas *et al.*, 2005):

$$\log OR_{DX} \approx \frac{\log OR_{DZ}}{\beta_{X|Z}}. \tag{6.9}$$

The above has the advantage that OR_{DZ} can be estimated from case-control data, even using a meta-analysis of several such case-control studies; while $\beta_{X|Z}$ can be estimated from

the controls only, or from separate studies. This estimator is called *Wald-OR* as such ratios have also been considered in the context of measurement error.

The advantages but also the shortcomings of estimation based on the Wald-OR (6.9) are discussed in Harbord *et al.* (2013). The main points are: it is not consistent either for a population COR or for a conditional COR, even when there is no unobserved confounding. Its justification is as approximation of the causal risk ratio for rare diseases under the assumption that the exposure X has a conditional normal distribution with a specific mean structure (Didelez *et al.*, 2010a). However, the estimator is consistent under the null because, using our earlier reasoning, we expect $OR_{DZ} = 1$ if and only if there is no causal effect. Moreover, Harbord *et al.* (2013) give a number of conditions, some but not all empirically testable, under which the asymptotic bias for estimating the marginal COR is smaller than 10%. In contrast, Didelez *et al.* (2010a) show for the case of a binary exposure, where the normality assumption clearly fails, that in many realistic scenarios the bias can easily be 30% or larger. So, unlike the TSLS-estimator, the performance of the Wald-OR is quite sensitive to the type of exposure distribution.

Using the Wald-OR (6.9) is especially popular in the context of Mendelian randomisation studies because the relations between disease and many genetic factors are easily available from summary data from case-control GWASs, and are in fact often combined in a meta-analytic way to obtain estimates of OR_{DZ}. Moreover, it does not require any assumptions such as knowledge of the disease prevalence. In fact, none of the alternative estimators (addressed below) can be computed when only summary data are available, or without knowledge of the disease prevalence.

Returning to the data from Meleady *et al.* (2003), we find that the log-OR between the genetic variant and CVD is 0.11 (se = 0.077); the linear regression coefficient from a regression of the binary exposure on the genetic factor within the controls is only 0.071 (se = 0.013) indicating a weak instrument; so a rough estimate according to (6.9) would be a COR of 4.73 (95% CI: (0.52; 43.3)). Note that if we used all data for the denominator, instead of only the controls, we would find a linear regression coefficient of 0.082 which would suggest a COR of 3.81 (95% CI: (0.58, 25.1)); the difference in (Z, X)-association for controls compared to the whole sample is due to the selection effect. The very wide confidence intervals are typical for a weak instrument.

6.6.3.2 Instrumental variable estimation using external information

In Mendelian randomisation studies, it is not unrealistic to assume that the population distribution of the gene alleles and hence $E(Z)$ is known. For a *multiplicative structural mean model (MSMM)* (Hernán and Robins, 2006) this is sufficient to estimate the causal risk ratio with an IV from case-control data as shown in Bowden and Vansteelandt (2011). A MSMM assumes

$$\frac{P(D = 1 \mid X, Z)}{P(D = 1 \mid X, Z; \mathrm{do}(X = 0))} = \exp(\psi X), \tag{6.10}$$

so the causal parameter of interest here is the logarithm of the risk ratio ψ, which describes the effect of exposure on the exposed. As an example consider again alcohol consumption: $\exp(\psi X)$ corresponds to the multiplicative reduction in risk of a person who by her own choice would drink an amount X of alcohol, but is forced by an intervention not to drink any.

The MSMM can be fitted using the following estimating equations which remain valid under case-control sampling

$$0 = \sum_i (Z_i - E(Z)) D_i \exp(-\psi X_i),$$

where we see the need to know $E(Z)$. If this is not available but the disease prevalence $\tilde{p}$ in the population is known, then we can instead use the estimate

$$\hat{E}(Z) = \frac{\sum W_i Z_i}{\sum W_i}$$

with weights $W_i = D_i\tilde{p}/\tilde{p}_S + (1 - D_i)(1 - \tilde{p})/(1 - \tilde{p}_S)$, and $\tilde{p}_S = P(D = 1 \mid S = 1)$. The solution to the resulting estimating equations, $\hat{\psi}$, is then a consistent estimator of the logarithm of the causal risk ratio, and its properties as well as formulas for standard errors are detailed in Bowden and Vansteelandt (2011). In contrast, if $\tilde{p}$ is unknown but assumed to be low, $E(Z)$ could be estimated from the controls only; the larger $\tilde{p}$ the more biased the estimator $\hat{\psi}$, though.

In analogy to (6.10), we could also assume a logistic SMM in order to estimate a causal OR describing the effect of exposure on the exposed (Vansteelandt and Goetghebeur, 2003). The estimating equations are more involved and require fitting an associational model for $P(D = 1 \mid X, Z)$ and we refer for the details again to Bowden and Vansteelandt (2011).

A philosophically different approach to combining an IV with case-control sampling has been proposed by Shinohara *et al.* (2012) who target the so-called *principal stratum* causal effect. This approach originates in the context of RCTs with partial compliance and estimates the 'effect of treatment on the compliers'. Again, IV-estimation of this type of causal effect relies on external information on the disease prevalence.

6.7 Conclusions

In conclusion we summarise the main issues of causal inference from case-control data (confounding, selection bias, retrospective sampling), and then discuss the implications for practical analyses.

6.7.1 Summary of main issues

With observational data, including case-control data, confounding poses the biggest threat to the validity of causal conclusions. We have given an intuitive and formal characterisation of when sufficient information to account for confounding is available; causal DAGs and the back-door criterion are useful tools to identify such information. When there is unobserved confounding, instrumental variables provide an alternative if available.

Another danger especially relevant to case-control sampling is selection bias. This can occur when the controls are not truly sampled from the population, but unrepresentative databases are used instead; recall bias and ascertainment bias also fall in this category. The null preservation property is relevant in this context: a method should at least give valid conclusions when the causal null hypothesis is true even if it is potentially biased otherwise; instrumental variables can be exploited in this way. We have also seen the duality between evaluating a causal null hypothesis and a causal odds ratio: selection bias is excluded from inference on both these targets by the same structural assumptions, which can easily be checked on causal DAGs. To construct the DAG and hence to justify these assumptions, a thorough understanding of the sampling process and its implications are crucial.

A third characteristic of case-control studies is that we observe $P(D, X, C \mid S = 1)$, and not the population distribution $P(D \mid X, C)$. However, many standard causal parameters are functions of the latter; marginal causal effects require some information on the covariate distribution $P(C)$ in the population. Such causal effect measures are often chosen with a view

to comparison with results from RCTs, as well as to public health decision making. However, we have seen that only conditional causal odds ratios can be obtained from case-control data without external information; further, their interpretation must take noncollapsibility into account. All other methods covered in this chapter require some additional information or are approximative. If certain independencies in the population are known, such as between genetic factors and exposure, these can help to reconstruct the marginal distribution. Knowledge of the disease prevalence $\tilde{p}$ enables standardisation, propensity score methods, IPTW, and IV methods; the distribution within controls can be used to approximate some of the ingredients for these methods when the disease is rare.

In view of the pivotal role of structural assumptions, as well as the possible need for external information which may often only be approximative, it appears that systematic sensitivity analyses are especially relevant to causal inference from case-control data. However, while sensitivity analyses have been considered in a few areas of causal inference, we are not aware of any such work for case-control designs.

6.7.2 Implication for practical analyses of case-control studies

Current practice in epidemiology when faced with case-control data is largely based on logistic regression modelling. As addressed in Section 6.5.1, this enables causal conclusions if the covariates included are sufficient to adjust for confounding, if a conditional causal odds-ratio is the target of inference, and if sources of selection bias, e.g., due to the available database, omitted covariates, recall or ascertainment bias, reverse causation, etc., can be eliminated. When aiming at reliable causal conclusions these assumptions should therefore carefully be addressed and justified, both by empirical evidence and subject matter knowledge; even if their full validity in a given application is doubtful, the source of possible violations and how these might be prevented in future studies or analyses should be clarified (and ideally supplemented by a sensitivity analysis).

We suggest that confidence in and understanding of – and hence critical discussion of – these crucial causal assumptions can always be strengthened by supplementing the analysis with a detailed graphical representation; this should contain the observed variables, relevant unobserved variables, design features, and the known or assumed structural relationships between these in the form of a DAG. While this allows us to graphically verify the assumptions, an important additional benefit of using DAGs is that structural assumptions are made *explicit and transparent* in the first place. When constructing the DAG for a given data situation it is important to justify the absence of further edges or nodes, as these would often result in a violation of assumption or a source of spurious association due to selection (such as if U in Figure 6.3(a) were ignored). Moreover, when addressing the problem of confounding, the user may find it easiest to think about the data generating mechanisms prospectively in order to obtain an appropriate DAG, with careful distinction between pre- and post-exposure covariates based on the way they are measured. Sampling selection would then be addressed by including further nodes and edges relevant to the sampling node S reflecting the study design, and conditioning on the latter. Examples for this were given with Figure 6.2(b) in Section 6.2 showing the problem of ascertainment bias (no adjustment with observed variables results in valid inference), and with Figure 6.5 in Section 6.3.3 illustrating a study on the effect of HRT on TCI (where more than one subset of the covariates allows valid adjustment).

Logistic regression further relies on correct model specification (except in the rare case when a saturated model can be used) and therefore does not satisfy the null-preservation property, but has the advantage of not requiring any extra outside information on disease prevalence. It is therefore important to remember that a valid test for the causal null hypothesis can be obtained under weaker assumptions, either based on Proposition 1 or using

an instrument as in Section 6.6.2, and we suggest that any analysis should be complemented by such a test.

In some applications, a logistic regression analysis may be considered too limiting, and this is where the other approaches presented in this chapter may be attractive. This is especially important when the user desires to partly relax some of the assumptions, or to estimate a causal parameter other than a conditional causal odds ratio for example, for better comparability with RCTs. While there are a few examples where alternatives to logistic regression have been applied (Persson and Waernbaum, 2013), more practical analyses are still required to gain a better understanding of their strength and weaknesses. In contrast, there are an increasing number of applications of IV methods for MR studies in case-control settings due to the availability of GWAS data (see references in Harbord *et al.* (2013)). Again, we recommend that such analyses be based on a transparent and well-justified DAG representation of the assumption. As before, this is crucial in order to exclude selection bias by design, or by conditioning on post-IV covariates (Aschard *et al.*, 2015); an explicit consideration of the causal null hypothesis using the IV-outcome association still relies on weaker assumptions, and we again suggest that any IV-analysis should be complemented by such a test.

6.8 Appendix: Directed acyclic graphs

Basic reading for anyone interested in (causal) DAGs are Spirtes *et al.* (2000), Pearl (2009), Dawid (2015), and Hernán and Robins (2018).

Terminology and basics

A *directed acyclic graph* (DAG) $G = (V, E)$ consists of a set of nodes (vertices) V and directed edges E between pairs of distinct nodes, where a single edge $(a, b) \in E$ is represented as $a \to b$. In a DAG there is only at most one type of edge between two nodes. An ordered sequence of distinct nodes $(v_1, \ldots, v_J)$ such that there is an edge in E between any two successive nodes is called a *path*, i.e., $(v_j, v_{j+1}) \in E$ or $(v_{j+1}, v_j) \in E$; if the edges all point in the same direction it is a *directed path*. A DAG is characterised by the assumption that there are no directed paths from one node to itself, i.e., no sequence $a \to \cdots \to a$. A back-door path from a to b is any path that starts with an edge pointing at a, i.e., with $a \leftarrow \cdots b$. Note that this definition is not symmetric in a, b.

If $a \to b$, then a is a *parent* of b, and the set of all *parents* is denoted by $\text{pa}(b)$. All nodes $b \in V \backslash \{a\}$ such that there is a directed path from a to b are called *descendants* of a; all others are its *nondescendants*. Each intermediate node $v_j, j = 2, \ldots, J-1$, on a path $(v_1, \ldots, v_J)$ is either a *collider*, if $\to v_j \leftarrow$, or a *noncollider* in all other cases. Note that the same node can be a collider on one path and a noncollider on another path.

A DAG is associated with a joint distribution of variables represented by the nodes through conditional independencies that correspond to absences of edges in the DAG *(Markov properties)*. First of all, we define that a joint distribution for $X_1 \ldots, X_K$ *factorises* according to a DAG $G = (V, E)$, $V = \{1 \ldots, K\}$ if the pdf/pmf satisfies

$$p(x_1, \ldots, x_K) = \prod_{i \in V} p(x_i | x_{\text{pa}(i)}).$$

Conditional independencies implied by this factorisation can be read off the DAG through 'd-separation', defined next. A path between two nodes a and b is *blocked* by a set $C \subset$

$V \backslash \{a, b\}$ if (i) the path contains a node $v \in C$ that is a noncollider; or (ii) it contains a collider v' such that neither v' nor any descendants of v' are elements of C. Then the sets A and B are *d-separated* by $C \subset V \backslash (A \cup B)$ if all paths between all $a \in A$ and $b \in B$ are blocked by C (see, e.g., Lauritzen (1996) for an equivalent way of reading off separations in a DAG using 'moralisation'). Now, for any distribution that factorises according to a DAG it holds that whenever A and B are d-separated by C then $X_A \perp\!\!\!\perp X_B \,|\, X_C$ (or we also write $A \perp\!\!\!\perp B \,|\, C$). Paths can be blocked by the empty set, e.g., if they contain a collider or if there is no path between two nodes in the first place.

Note that while the factorisation implies that any d-separation corresponds to a conditional independence, the converse is not true. It is possible for a distribution to contain additional conditional or marginal independencies that cannot be seen from the DAG, e.g., due to 'cancellation' of paths. A joint distribution where this does not happen is called *faithful* to the DAG.

Causal DAGs and back-door criterion

A DAG is supplemented with a causal interpretation by demanding that the parents $X_{\mathrm{pa}(a)}$ of a node/variable X_a are its *direct causes* (relative to the given set of nodes/variables X_V). This 'direct cause' requirement means that if we fix $X_{\mathrm{pa}(a)}$ by intervention, manipulations of any other nodes do not have an effect on X_a, while manipulations of any of the parent nodes individually (while holding the other ones fixed) have an effect on X_a (Spirtes *et al.*, 2000). A further requirement is that any 'common cause' of any two variables X_a, X_b must be included in the DAG whether observable or not. While any conditional independencies implied by the DAG involving observable variables can (and should) be checked empirically, these requirements essentially need to be justified based on subject matter knowledge. For example, whether a given set C is sufficient to adjust for confounding, or the IV conditions cannot be (entirely) verified empirically from observational data.

The above definition of a causal DAG can be somewhat relaxed but we will not go into details here. A number of alternative causal interpretations supplementing DAGs with additional or different types of nodes have been proposed (Dawid, 2002; Richardson and Robins, 2013). Another modification common in much of the causality literature is not to represent unobservable variables explicitly as nodes, but to use bi-directed edges $a \leftrightarrow b$ when $a \leftarrow U \rightarrow b$ for an unobservable U; again we do not go into details here (Pearl, 2009).

A popular criterion that is used to find, from a DAG, a set of covariates that is sufficient to adjust for confounding is the *back-door criterion* (Pearl, 1995). This states that given a causal DAG on a set of variables V, we can identify the causal effect of $X \in V$ on $D \in V$ if a set $C \subset V \backslash \{X, D\}$ exists such that C are nondescendants of X and C blocks every back-door path from X to D. The causal effect is then given by (6.6).

Causal DAGs can further be used to check other criteria under which causal effects are identified, leading to different formulae. In fact, a complete characterisation of identification of causal effects can be given. However, this complete characterisation does not cover the case of identification under sampling selection (Bareinboim and Tian, 2015; Bareinboim *et al.*, 2014).

Bibliography

Aschard, H., Vilhjálmsson, B. J., Joshi, A. D., Price, A. L., and Kraft, P. (2015). Adjusting for heritable covariates can bias effect estimates in genome-wide association studies. *The*

American Journal of Human Genetics, **96**, 329 – 339.

Balke, A. and Pearl, J. (1994). Counterfactual probabilities: Computational methods, bounds, and applications. In *Proceedings of the Tenth Annual Conference on Uncertainty in Artificial Intelligence*, pages 46–54. Morgan Kauffman, San Francisco.

Bareinboim, E. and Pearl, J. (2012). Controlling selection bias in causal inference. In *Proceedings of the Fifteenth International Conference on Artificial Intelligence and Statistics*, volume 22 of *Proceedings of Machine Learning Research*, pages 100–108.

Bareinboim, E. and Tian, J. (2015). Recovering causal effects from selection bias. In *Proceedings of the Twenty-Ninth National Conference on Artificial Intelligence*, pages 3475–3481. AAAI Press, Menlo Park, CA.

Bareinboim, E., Tian, J., and Pearl, J. (2014). Recovering from selection bias in causal and statistical inference. In *Proceedings of the 28th National Conference on Artificial Intelligence*, pages 2410–2416. AAAI Press, Menlo Park, CA.

Berkson, J. (1946). Limitations of the application of fourfold table analysis to hospital data. *Biometrics Bulletin*, **2**, 47–53.

Borboudakis, G. and Tsamardinos, I. (2015). Bayesian network learning with discrete case-control data. In *Proceedings of the 31st Conference on Uncertainty in Artificial Intelligence*, pages 151–160. AUAI Press, Arlington, Virginia.

Bowden, J. and Vansteelandt, S. (2011). Mendelian randomization analysis of case-control data using structural mean models. *Statistics in Medicine*, **30**, 678–694.

Casas, J. P., Bautista, L. E., Smeeth, L., Sharma, P., and Hingorani, A. D. (2005). Homocysteine and stroke: Evidence on a causal link from Mendelian randomisation. *The Lancet*, **365**, 224–232.

Cooper, G. F. (1995). Causal discovery from data in the presence of selection bias. In *Proceedings of the Fifth International Workshop on Artificial Intelligence and Statistics*, pages 140–150. Morgan Kaufmann, San Francisco.

Davey Smith, G. and Ebrahim, S. (2003). Mendelian randomization: Can genetic epidemiology contribute to understanding environmental determinants of disease? *International Journal of Epidemiology*, **32**, 1–22.

Dawid, A. P. (1979). Conditional independence in statistical theory (with discussion). *Journal of the Royal Statististical Society, Series B*, **41**, 1–31.

Dawid, A. P. (2000). Causal inference without counterfactuals (with discussion). *Journal of the American Statistical Association*, **95**, 407–448.

Dawid, A. P. (2002). Influence diagrams for causal modelling and inference. *International Statistical Review*, **70**, 161–89.

Dawid, A. P. (2015). Statistical causality from a decision-theoretic perspective. *Annual Review of Statistics and Its Application*, **2**, 273–303.

Didelez, V. and Sheehan, N. A. (2007a). Mendelian randomisation as an instrumental variable approach to causal inference. *Statistical Methods in Medical Research*, **16**, 309–330.

Didelez, V. and Sheehan, N. A. (2007b). Mendelian randomisation: Why epidemiology needs a formal language for causality. In F. Russo and J. Williamson, editors, *Causality and Probability in the Sciences*, volume 5 of *Texts in Philosophy*, pages 263–292. College Publications, London.

Didelez, V., Meng, S., and Sheehan, N. A. (2010a). Assumptions of IV methods for observational epidemiology. *Statistical Science*, **25**, 22–40.

Didelez, V., Kreiner, S., and Keiding, N. (2010b). Graphical models for inference under outcome-dependent sampling. *Statistical Science*, **25**, 368–387.

Edwards, D. and Lauritzen, S. L. (2001). The TM algorithm for maximising a conditional likelihood function. *Biometrika*, **88**, 961–972.

Evans, R. J. and Didelez, V. (2015). Recovering from selection bias using marginal structure in discrete models. In *Proceedings of the 31st Annual Conference on Uncertainty in Artifical Intelligence – Causality Workshop*, pages 46–55. AUAI Press, Corvallis, Oregon.

Goldberger, A. (1972). Structural equation methods in the social sciences. *Econometrica*, **40**, 979–1001.

Greenland, S. (2003). Quantifying biases in causal models: Classical confounding vs collider-stratification bias. *Epidemiology*, **14**, 300–306.

Greenland, S., Robins, J. M., and Pearl, J. (1999). Confounding and collapsibility in causal inference. *Statistical Science*, **14**, 29–46.

Harbord, R. M., Didelez, V., Palmer, T. M., Meng, S., Sterne, J. A. C., and Sheehan, N. A. (2013). Severity of bias of a simple estimator of the causal odds ratio in Mendelian randomization studies. *Statistics in Medicine*, **32**, 1246–1258.

Hernán, M. A. (2004). A definition of causal effect for epidemiological research. *Journal of Epidemiology and Community Health*, **58**, 265–271.

Hernán, M. A. and Robins, J. M. (2006). Instruments for causal inference: An epidemiologist's dream? *Epidemiology*, **17**, 360–372.

Hernán, M. A. and Robins, J. M. (2018). *Causal Inference*. Chapman & Hall/CRC, Boca Raton. Forthcoming.

Hernán, M. A., Hernández-Díaz, S., and Robins, J. M. (2004). A structural approach to selection bias. *Epidemiology*, **15**, 615–625.

Hill, A. B. (2015). The environment and disease: Association or causation? *Journal of the Royal Society of Medicine*, **108**, 32–37.

Ioannidis, J. P. A. (2016). Exposure-wide epidemiology: Revisiting Bradford Hill. *Statistics in Medicine*, **35**, 1749–1762.

Keiding, N. and Clayton, D. (2014). Standardization and control for confounding in observational studies: A historical perspective. *Statistical Science*, **29**, 529–558.

Lauritzen, S. L. (1996). *Graphical Models*. Clarendon Press, Oxford.

Månsson, R., Joffe, M. M., Sun, W., and Hennessy, S. (2007). On the estimation and use of propensity scores in case-control and case-cohort studies. *American Journal of Epidemiology*, **166**, 332–339.

Meleady, R., Ueland, P. M., Blom, H., Whitehead, A. S., Refsum, H., Daly, L. E., Vollset, S. E., Donohue, C., Giesendorf, B., Graham, I. M., Ulvik, A., Zhang, Y., Bjorke Monsen, A.-L., and the EC Concerted Action Project: Homocysteine and Vascular Disease (2003). Thermolabile methylenetetrahydrofolate reductase, homocysteine, and cardiovascular disease risk: The European Concerted Action Project. *The American Journal of Clinical Nutrition*, **77**, 63–70.

Moerkerke, B., Vansteelandt, S., and Lange, C. (2010). A doubly robust test for gene-environment interaction in family-based studies of affected offspring. *Biostatistics*, **11**, 213–225.

Palmer, T. M., Ramsahai, R. R., Didelez, V., and Sheehan, N. A. (2011). Nonparametric bounds for the causal effect in a binary instrumental-variable model. *Stata Journal*, **11**, 345–367.

Pearl, J. (1995). Causal diagrams for empirical research. *Biometrika*, **82**, 669–688.

Pearl, J. (2009). *Causality*. Cambridge University Press, second edition.

Pedersen, A. T., Lidegaard, O., Kreiner, S., and Ottesen, B. (1997). Hormone replacement therapy and risk of non–fatal stroke. *The Lancet*, **350**, 1277–1283.

Persson, E. and Waernbaum, I. (2013). Estimating a marginal causal odds ratio in a case-control design: Analyzing the effect of low birth weight on the risk of type 1 diabetes mellitus. *Statistics in Medicine*, **32**, 2500–2512.

Prentice, R. L. and Pyke, R. (1979). Logistic disease incidence models and case-control studies. *Biometrika*, **66**, 403–411.

Richardson, T. S. and Robins, J. M. (2013). Single world intervention graphs (SWIGs): A unification of the counterfactual and graphical approaches to causality. *Working Paper No.128*. Center for Statistics and the Social Sciences of the University of Washington.

Robins, J. M. (2001). Data, design and background knowledge in etiologic inference. *Epidemiology*, **11**, 313–320.

Robins, J. M., Hernán, M. A., and Brumback, B. (2000). Marginal structural models and causal inference in epidemiology. *Epidemiology*, **11**, 550–560.

Rosenbaum, P. R. and Rubin, D. B. (1983). The central rôle of the propensity score in observational studies for causal effects. *Biometrika*, **70**, 41–55.

Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology*, **66**, 688–701.

Shinohara, R. T., Frangakis, C. E., Platz, E., and Tsilidis, K. (2012). Designs combining instrumental variables with case-control: Estimating principal strata causal effects. *The International Journal of Biostatistics*, **8**, 1–21.

Spirtes, P., Glymour, C., and Scheines, R. (2000). *Causation, Prediction and Search*. MIT Press, Cambridge, Massachusetts, second edition.

van der Laan, M. J. (2008). Estimation based on case-control designs with known prevalence probability. *The International Journal of Biostatistics*, **4**, Issue 1, Article 17.

VanderWeele, T. J. (2015). *Explanation in Causal Inference: Methods for Mediation and Interaction*. Oxford University Press.

VanderWeele, T. J. and Shpitser, I. (2013). On the definition of a confounder. *The Annals of Statistics*, **41**, 196–220.

Vansteelandt, S. and Goetghebeur, E. (2003). Causal inference with generalized structural mean models. *Journal of the Royal Statistical Society, Series B*, **65**, 817–835.

Whittemore, A. S. (1978). Collapsibility of multidimensional contingency tables. *Journal of the Royal Statististical Society, Series B*, **40**, 328–340.

Windmeijer, F. and Didelez, V. (2018). IV methods for binary outcomes. In G. Davey Smith, editor, *Mendelian Randomization: How genes can reveal the biological and environmental causes of disease*. Oxford University Press. Forthcoming.

Wooldridge, J. M. (2010). *Econometric Analysis of Cross Section and Panel Data*. MIT Press, Cambridge, Massachusetts.

Zeger, S. L., Liang, K.-Y., and Albert, P. S. (1988). Models for longitudinal data: A generalized estimating equation approach. *Biometrics*, **44**, 1049–1060.