1

Statistical causality: Some historical remarks

D.R. Cox
Nuffield College, University of Oxford, UK

1.1 Introduction

Some investigations are essentially descriptive. Others are concerned at least in part with probing the nature of dependences.

Examples of the former are studies to estimate the number of whales in a portion of ocean, to determine the distribution of particle size in a river bed and to find the mortality rates of smokers and of nonsmokers. Examples of the second type of investigation are experiments to compare the effect of different levels of fertilizer on agricultural yield and investigations aimed to understand any apparent differences between the health of smokers and nonsmokers, that is to study whether smoking is the explanation of differences found. Also much, but not all, laboratory work in the natural sciences comes in this category.

Briefly the objectives of the two types are respectively to describe the world and in some sense to understand it. Put slightly more explicitly, in the agricultural field trial the object is essentially to understand how the yield of a plot would differ if this level of fertilizer were used rather than that level or, in smoking studies, how the outcomes of subjects who smoke compare with what the outcomes would have been had they not smoked. These are in some sense studies of causality, even though that word seems to be sparingly used by natural scientists and until recently by statisticians.

Neyman (1923) defined a basis for causal interpretation, using a working and not directly testable assumption of unit-treatment additivity specifying the way a particular experimental unit would respond to various treatments, only one of which could actually be used for a specific unit. In the absence of randomization specific consequences were unclear. His

Causality: Statistical Perspectives and Applications, First Edition. Edited by Carlo Berzuini, Philip Dawid and Luisa Bernardinelli.
 Published 2012 by John Wiley & Sons, Ltd.

later more ambitious work (Neyman *et al*., 1935) is discussed below. The landmark paper on observational studies by Cochran (1965) did deal with causal interpretation and moreover in the discussion Bradford Hill outlined his considerations pointing towards stronger interpretation. Both authors emphasized the difficulties of interpretation in experiments and in observational contexts.

1.2 Key issues

Key issues are first what depth of understanding is involved in claims of causality, that is just what is meant by such a claim. Then there is the crucial matter of the security with which such causality can be established in any specific context. It is clear that defining models aimed at causal interpretation is desirable, but that calling such models causal and finding a good empirical fit to data is in some specific instances far short of establishing causality in a meaningful subject-matter sense. Cox and Wermuth (1996, pp. 219–) suggested *potentially causal* as a general term for such situations. This was criticized as overcautious by Pearl (2000). Indeed, overcaution in research is in general not a good idea. In the present situation, however, especially in a health-related context, caution surely is desirable in the light of the stream of information currently appearing, suggesting supposedly causal and often contradictory interpretations about diet and so forth.

1.3 Rothamsted view

The key principles of experimental design, developed largely at Rothamsted (Fisher, 1926, 1935; Yates, 1938, 1951), contain the essential elements of what is often needed to achieve reasonably secure causal conclusions in an experimental context, even though, as far as I can recall, the word cause itself is rarely if ever used in that period. An attempt to review that work largely in nontechnical terms (Cox, 1958a) centred the discussion on:

- a distinction between on the one hand experimental units and their properties and on the other hand treatments;
- a working assumption of unit-treatment additivity;
- a general restriction of supposedly precision-enhancing adjustments to features measured before randomization;
- a sevenfold classification of types of measurement with the above objectives in mind.

In more modern terminology the third of these points requires conditioning on features prior to the decision on treatment allocation and marginalization over features between treatment and final outcome. It presupposes, for example, the absence of what in a clinical trial context is called *noncompliance* or *nonadherence*. More generally, any intervention in the system between treatment allocation and response should either be independent of the treatment or be reasonably defined as an intrinsic part of the treatment. A possibly apocryphal agricultural example is that of a particular fertilizer combination so successful that the luxuriant growth of crop on the relevant plots attracted birds from afar who obtained their food solely from those plots. Consequently, these plots ended up showing a greatly depressed yield. Thus in

a so-called intention-to-treat analysis the treatment was a failure, a conclusion correct in one sense but misleading both scientifically and in terms of practical implication.

If some measurements had been available plot by plot on the activity of the birds, the situation might have been partly rescued, although by making strong assumptions. In general, causal interpretation of randomized experimental designs may be inhibited by unobserved, uncontrolled and unwanted interventions in the period between implementing the randomization and the measurement of outcome. This consideration is of especial concern if there is an appreciable time gap between randomization and measurement of outcome.

Causal interpretation of observational studies, on the other hand, is handicapped primarily by the possibility of unobserved confounders, that is unobserved explanatory variables. When the effect measured as a relative risk is large some protection against the distortions induced by such unobserved variables is provided by Cornfield's inequality (Cornfield *et al*., 1959).

1.4 An earlier controversy and its implications

Neyman and colleagues read to the Royal Statistical Society (Neyman *et al*., 1935) an account of work in which the notion of unit-treatment additivity, stemming from Neyman (1923) and in a sense implicitly underpinning the Rothamsted work on design, was replaced by a much more general assumption in which different experimental units had different treatment effects. A provisional conclusion of the analysis was that the standard estimate of error for a Latin square design is inappropriate. This suggestion led to a vehement denunciation by R.A. Fisher, recorded in appreciable detail in the published version of the discussion. The general issues of the role of randomization were further discussed in the next few years, mostly in *Biometrika*, with contributions from Student, Yates, Neyman and Pearson, and Jeffreys. With the exception of Student's contribution, which emphasized the role of randomization in escaping biases arising from personal judgement, the discussion focused largely on error estimation. Interestingly, the aspect most stressed in current writing, namely the decoupling from the estimation of treatment effects of unobserved systematic explanatory features, is not emphasized.

The specific issue of the Latin square was developed in more detail nearly 20 years later by Kempthorne and his colleagues, confirming (Kempthorne and Wilk, 1957) the bias of the error estimate in the Latin square analysis. Cox (1958b) pointed out the unreality of the null hypothesis being tested, namely that in spite of unexplained variation in treatment effect the null hypothesis considered was that the treatment effects balanced out *exactly* over the finite set of units in a trial. When a more realistic null hypothesis was formulated the biases disappeared, restoring the respectability of the Latin square analysis, as argued by Fisher.

While the status of error estimation in the Latin square may seem a rather specialized or even esoteric matter in the present context, the relevance for current discussions is this. It is common to define an average causal effect in essentially the way that Neyman *et al*. did, that is without any assumption that the effect is identical for all units of study. Of course, if the variation in treatment effect between units of study can be explained in substantive terms that is preferable, but if not, it should in some sense be treated as stochastic, and that affects the assessment of error.

The point is similar in spirit to what is sometimes called the marginalization principle, namely that it rarely makes sense to consider models with nonzero interaction but exactly zero main effects.

1.5 Three versions of causality

In general terms, it is helpful to distinguish three types of statistical causality, all of importance in appropriate contexts (Cox and Wermuth, 2004). The first is that essentially a multiple-regression like analysis shows a dependence not explained away by other appropriate explanatory variables. In a time series context this forms Wiener–Granger causality; see, also, the more general formulation by Schweder (1970). The second definition is in the spirit of the previous section in terms of a real or notional intervention and implies a restriction on the nature of variables to be treated as possibly causal. This is the approach that has received most attention in recent years (Holland, 1986). The third notion requires some evidence-based understanding in terms of the underlying process.

The general principle that causality operates in time and rarely instantaneously has implications for model formulation. Thus if two variables measured as discrete-time series (X_t, Y_t) are such that each component in some sense causes the other a suitable formulation is that (X_{t+1}, Y_{t+1}) depends on (X_t, Y_t), with a corresponding differential equation form in continuous time. There are appreciable difficulties for subjects like macro-economics, in which data are typically quite heavily aggregated in time (Hoover, 2001).

1.6 Conclusion

It seems clear that the objective of many lines of research is to establish dependencies that are in some sense causal. Models and associated methods of statistical analysis that point towards causality are therefore very appealing and important. However, the circumstances under which causality in a meaningful sense can be inferred from a single study may be relatively restricted, mainly to randomized experiments with clear effects and no possibility of appreciable noncompliance and sometimes to observational studies in which large relative risks are encountered. Really secure establishment of causality is most likely to emerge from qualitative synthesis of the conclusions from different kinds of study. A fine example is the paper on smoking and lung cancer by Cornfield *et al*. (1959), reprinted in 2009. Here large-scale population data, the outcomes of longitudinal prospective and retrospective studies and the results of laboratory work were brought together to provide a powerful assertion of causal effect in the face of some scepticism at the time.

References

Cochran, W.G. (1965) The planning of human studies of human populations (with discussion). *Journal of the Royal Statistical Socity science* A, **128**, 234–266.

Cornfield, J., Haenszel, W., Hammond, E.C., Lilienfeld, A.M., Shimkin, M.B. and Wynder, E.L. (1959) Smoking and lung cancer: recent evidence and a discussion of some questions. *Journal of the National Cancer Institute* **22**, 173–203. Reprinted 2009 in *International Journal of Epidemiology*, **38**, 1175–1191.

Cox, D.R. (1958a) *Planning of Experiments*. New York: John Wiley & Sons, Inc.

Cox, D.R. (1958b) The interpretation of the effects of non-additivity in the Latin square. *Biometrika*, **45**, 69–73.

Cox, D.R. and Wermuth, N. (1996) *Multivariate Dependencies*. London: Chapman & Hall.

Cox, D.R. and Wermuth, N. (2004) Causality: a statistical view. *International Statistical Review*, **72**, 285–305.

Fisher, R.A. (1926) The arrangement of field experiments. *Journal of the Ministry of Agriculture Great Britain*, **33**, 503–513.

Fisher, R.A. (1935) *Design of Experiments*. Edinburgh: Oliver and Boyd, and subsequent editions.

Holland, P.W. (1986) Statistics and causal inference (with discussion). *Journal of the American Statistical Association*, **81**, 945–970.

Hoover, K.V. (2001) *Causality in Macroeconomics*. Cambridge: Cambridge University Press.

Kempthorne, O. and Wilk, M.B. (1957) Nonadditivity in a Latin square design. *Journal of the American Statistical Association*, **52**, 218–236.

Neyman, J. (1923) On the application of probability theory to agricultural experiments. Essay on principles. *Rocniki Nauk Rolniiczych*, **10**, 1–51. English translation of Section 9 by D.M. Dabrowska and T.P. Speed, *Statistical Science*, **9**, 465–480.

Neyman, J., Iwaszkiewicz, K. and Kolodziejczyk, St. (1935) Statistical problems in agricultural experimentation (with discussion). *Supplement of the Journal of the Royal Statistical Society*, **2**, 107–180.

Pearl, J. (2000) *Causality*. Cambridge: Cambridge University Press, 2nd edn, 2010.

Schweder, T. (1970) Composable Markov processes. *Journal Applied Probability*, **7**, 400–410.

Yates, F. (1938) *Design and Analysis of Factorial Experiments*. Harpenden: Imperial Bureau of Soil Science.

Yates, F. (1951) Bases logiques de la planification des experiences. *Annals of the Institute of H. Poincaré*, **12**, 97–112.

2

The language of potential outcomes

Arvid Sjölander
Department of Medical Epidemiology and Biostatistics, Karolinska Institutet, Stockholm, Sweden

2.1 Introduction

A common aim of empirical research is to study the association between a particular exposure and a particular outcome. In epidemiology, for instance, many studies have been conducted to explore the association between smoking and lung cancer. The ultimate goal, however, is often more ambitious; typically we want to establish a causal relationship, e.g. that smoking actually causes lung cancer. It is well known that association does not imply causation; an observed association between smoking and lung cancer may have several explanations, which do not necessarily involve a causal smoking effect. Thus, an important question for empirical researchers is this:

> Under what conditions can an observed association be interpreted as a causal effect?

This question is logically precluded by the following:

> What do we mean, more precisely, when we talk about causal effects?

Despite their obvious relevance, these questions were not treated formally in the statistical literature until the late 1970s. The dominating paradigm was that 'statistics can only tell us about association and not causation'. Thus, for most of the 20th century, causality remained an

Causality: Statistical Perspectives and Applications, First Edition. Edited by Carlo Berzuini, Philip Dawid and Luisa Bernardinelli.
 Published 2012 by John Wiley & Sons, Ltd.

ill-defined concept, and empirical researchers who wanted to draw causal conclusions from data had to resort to informal reasoning and justification. During the end of the century, however, a formal theory of causal inference emerged, based on *potential outcomes*. The foundation was laid out by Rubin (1974, 1978, 1980) for point exposures, that is exposures which are measured at a single point in time. Robins (1986, 1989) extended the potential outcome framework to time-varying exposures. Pearl (1993) demonstrated that important conceptual insights can be made by interpreting the potential outcomes as solutions to a nonparametric structural equation (NPSE) system. Galles and Pearl (1998) provided a complete axiomatization of the framework. In this chapter we provide a brief summary of the central ideas in the potential outcomes framework, as formulated by Rubin (1974, 1978, 1980). Specifically, we demonstrate in Sections 2.2 and 2.3 how the two questions above can be formally phrased and answered within this framework.

2.2 Definition of causal effects through potential outcomes

2.2.1 Subject-specific causal effects

Suppose that a particular subject dies after a surgical intervention. Did the intervention cause the subject's death? A natural way to answer this question would be to try to imagine what would have happened to this subject had he not been operated on. If the subject would have died anyway, we would not say that intervention was the cause. If, on the other hand, the subject would not have died had he not been operated on, then we would say that intervention caused the death. Thus, to determine whether an exposure causes an outcome we typically make a mental comparison between two scenarios; one where the exposure is present and one were the exposure is absent. If the outcome differs between the two scenarios we say that exposure has a causal effect on the outcome. The potential outcomes framework formalizes this intuitive approach to causality.

Specifically, suppose that we want to learn about the causal effect of a particular point exposure on a particular outcome. Let Y_x denote the outcome for a randomly selected subject in the study population, if the subject would hypothetically receive exposure level x. Depending on what exposure level the subject actually receives, Y_x may or may not be realized. For this reason, Y_x is referred to as a *potential* outcome. Let $\mathbb{X}$ be the set of all possible exposure levels. Each subject is coupled with a set of potential outcomes, $\{Y_x\}_{x\in\mathbb{X}}$, which completely describes how the outcome would look like for the particular subject, under each possible level of the exposure. For instance, when the exposure is binary we may define $\mathbb{X} = \{0, 1\}$, with 0 and 1 corresponding to 'unexposed' and 'exposed', respectively. In this case, each subject is coupled with a potential outcome vector $\{Y_0, Y_1\}$, where Y_0 is the outcome if the subject is unexposed and Y_1 is the outcome if the subject is exposed.

In practice, the outcome for a given subject may depend on the exposure and/or outcome for other subjects in the population. This would, for example, be the case in a vaccine intervention trial for an infectious disease, where the infection status of one subject depends on whether other subjects are infected as well. When subjects interfere in this way, a causal analysis based on potential outcomes becomes more involved. We will proceed by assuming that the outcome for a given subject does not depend on the exposure or outcome for other subjects. This assumption is a part of the 'stable-unit-treatment-value-assumption (SUTVA)' by Rubin

(1980). See Hudgens and Halloran (2008) and Tchetgen Tchetgen and VanderWeele (2010) for extensions to scenarios with subject interference.

Using potential outcomes, we define the *subject-specific causal effect* of exposure level x' versus x'' as some contrast of the potential outcomes $Y_{x'}$ and $Y_{x''}$, for example $Y_{x'} - Y_{x''}$. When the exposure is binary, the subject-specific causal effect is a contrast between Y_0 and Y_1. If, for a given subject, all potential outcomes are equal (i.e. Y_x does not depend on x), then, for this subject, the exposure has no causal effect on the outcome. If the exposure has no causal effect on the outcome for any subject in the study population, then we say that the *sharp causal null hypothesis* holds. A fundamental problem with subject-specific causal effects is that they are notoriously difficult to identify. This is because we cannot in general observe the same subject under several exposure levels simultaneously. Let X and Y denote the observed exposure and outcome, respectively, for a given subject. If a subject is exposed to level $X = x'$, then the potential outcome $Y_{x'}$ is assumed to be equal to the observed factual outcome Y for that subject. This link between the potential outcomes and the factual outcome is usually referred to as the 'consistency assumption', and is formally expressed as

$$X = x' \Rightarrow Y_{x'} = Y \tag{2.1}$$

We remain ignorant, however, about would have happened to the subject had it been exposed to some other level. Thus, for a subject who is exposed to level $X = x'$, all potential outcomes in $\{Y_x\}_{x \in \mathbb{X}}$, except $Y_{x'}$, are unobserved, or *counterfactual*. The word 'counterfactual' echoes the fact that the unobserved potential outcomes correspond to scenarios which are 'contrary to fact' – they did not happen. When the exposure is binary, Y_0 is unobserved if the subject is exposed and Y_1 is unobserved if the subject is unexposed. Thus, constrasts between Y_0 and Y_1 cannot be observed for any subject. Because subject-specific causal effects are in general not identifiable, they are of limited practical value.[1]

2.2.2 Population causal effects

A more useful concept is the *population causal effect*, which measures the aggregated impact of the exposure over the study population. Because the potential outcome Y_x may vary across subjects, we may treat it as a random variable, following a probability distribution $\Pr(Y_x)$ (we use $\Pr(\cdot)$ generically for both probabilities and densities). We interpret $\Pr(Y_x = y)$ as the population proportion of subjects with an outcome equal to y under the hypothetical scenario where everybody receives exposure level x. The population causal effect of exposure level x' versus x'' is defined as a contrast between the potential outcome distributions $\Pr(Y_{x'})$ and $\Pr(Y_{x''})$, for example the causal mean difference $E(Y_{x'}) - E(Y_{x''})$. When the outcome Y is binary, it would be natural to consider the causal risk ratio $\Pr(Y_x = 1)/\Pr(Y_{x'} = 1)$ or the causal odds ratio $[\Pr(Y_x = 1)/\Pr(Y_x = 0)]/[\Pr(Y_{x'} = 1)/\Pr(Y_{x'} = 0)]$. If $\Pr(Y_x)$ does not depend on x, then the exposure has no population causal effect on the outcome; we say that the causal null hypothesis holds. The sharp causal null hypothesis implies the causal null hypothesis. The converse is not true though; it is logically possible that the exposure has a causal effect for some subjects, but that these effects 'cancel out' in such a way that there is no aggregated effect over the population. In Section 2.3 we demonstrate that the population causal effect can

[1] A rare exception is when we are able to observe the same subject under several exposure levels subsequently, without any crossover effects. In these situations, subject-specific causal effects can be identified.

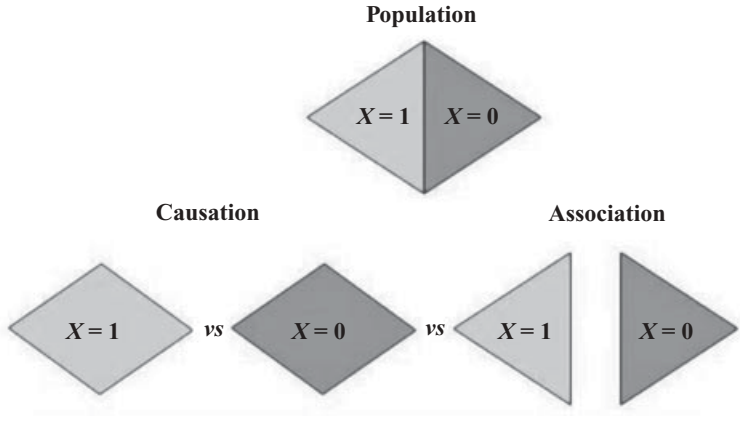

Figure 2.1 Association versus causation.

be identified even if we are not able to observe the same subject under several exposure levels simultaneously.

2.2.3 Association *versus* causation

Using potential potential outcomes, the fundamental difference between association and causation becomes clear. In the population of Figure 2.1 (adopted from Hernán and Robins forthcoming), some subjects are exposed ($X = 1$) and some subjects are unexposed ($X = 0$). We say that exposure and outcome are associated in the population if the outcome distribution differs between the exposed and unexposed. To quantify the association we may, for instance, use the mean difference $E(Y|A = 1) - E(Y|A = 0)$ or the risk ratio $\Pr(Y = 1|A = 1)/\Pr(Y = 1|A = 0)$. Thus, when we assess the exposure–outcome association we are by definition comparing two *different groups of subjects*: those who are factually exposed against those who are factually unexposed. In contrast, the population causal effect compares the potential outcomes for *the same* subjects (the whole population) under two hypothetical scenarios: everybody being exposed versus everybody being unexposed. This fundamental difference is the reason why association is in general not equal to causation. When we compare different subjects, there is always a risk that the subjects are different in other aspects than in the received exposure levels. If they are, then we may observe different outcome distributions for the exposed and the nonexposed, even if the exposure has no causal effect on the outcome.

2.3 Identification of population causal effects

2.3.1 Randomized experiments

As discussed in the previous section, population causal effects compare the outcome distributions for the *same* population under two different exposure levels. In practice, however, we cannot observe the same population under several exposure levels simultaneously. Nevertheless, intuition tells us that observed associations (i.e. comparisons between *different* subjects)

can be interpreted as causal effects under certain conditions. Specifically, it is widely accepted that 'association equals causation' in well-designed randomized experiments. Using potential outcomes we can give formal support to this notion. If the exposure is properly randomized, then all pre-randomization variables (i.e. variables whose values are determined prior to randomization, e.g. age, sex, eye color) are equally distributed, asymptotically, across levels of the exposure. Now, note that the potential outcomes $\{Y_x\}_{x\in\mathbb{X}}$ are by definition pre-randomization variables; they describe how the given subject will react to different exposure levels, which depends on intrinsic subject characteristics determined prior to randomization. Thus, in a randomized experiment, the potential outcomes are statistically independent of the exposure. We express this formally as

$$\{Y_x\}_{x\in\mathbb{X}} \amalg X \tag{2.2}$$

When (2.2) holds, subjects are said to be *exchangeable* across exposure levels.[2] Under consistency (2.1) and exchangeability (2.2), the conditional probability of Y, among those who actually received exposure level x, is equal to the probability of Y, had everybody received level x:

$$\Pr(Y = y|X = x) = \Pr(Y_x = y|X = x) = \Pr(Y_x = y) \tag{2.3}$$

The first equality in (2.3) follows from (2.1) and the second equality follows from (2.2). Thus, under consistency and exchangeability, any measure of association between X and Y equals the 'corresponding' population causal effect of X on Y. For instance, the associational mean difference $E(Y|A = 1) - E(Y|A = 0)$ equals the causal mean difference $E(Y_1) - E(Y_0)$ and the associational relative risk $\Pr(Y = 1|X = 1)/\Pr(Y = 1|X = 0)$ equals the causal risk ratio $\Pr(Y_1 = 1)/\Pr(Y_0 = 1)$. Because randomization produces exchangeability it follows that population causal effects are identifiable in randomized experiments.

We make a technical remark. Exchangeability, as defined in (2.2), means that all potential outcomes in $\{Y_x\}_{x\in\mathbb{X}}$ are *jointly* independent of X. Although this is a sufficient criterion for identification of the population causal effects, it is slightly stronger than necessary. By inspecting (2.3) we observe that $\Pr(Y_x)$ is identified for all x if the potential outcomes in $\{Y_x\}_{x\in\mathbb{X}}$ are *separately* independent of X:

$$Y_x \amalg X \qquad \forall x \tag{2.4}$$

In the literature, the word 'exchangeability' is sometimes used for the relation in (2.4). Rosenbaum and Rubin (1983) referred to (2.2) and (2.4) as 'strongly ignorable treatment assignment' and 'ignorable treatment assignment', respectively. Strong ignorability implies ignorability, but not the other way around. It is difficult, however, to imagine a practical scenario where the latter holds but not the former.

[2] We note that this definition of 'exchangeable' is different from the definition used in other branches of statistics, e.g. de Finetti (1974).

2.3.2 Observational studies

For ethical or practical reasons, randomization of the exposure is not always feasible. Indeed, many scientific 'truths' have been established through observational (i.e. nonrandomized) studies. When the exposure is not randomized, exchangeability does not necessarily hold, and an observed association cannot in general be interpreted as a causal effect. Violations of (2.2) typically occur when the exposure and the outcome have common causes. As an illustration, suppose that we wish to study the effect of a medical treatment (X) on the prognosis (Y) for patients with a certain disease. Suppose that a patients's general health status affects what treatment level the patient is assigned to (patients in a critical condition may, for example, receive higher doses than patients in a noncritical condition). Moreover, a patient's health status clearly affects his/her prognosis. That a patients's health status affects both treatment and prognosis implies that X and $\{Y_x\}_{x\in\mathbb{X}}$ are associated, which violates (2.2). When the exposure and outcome have common causes we say that the exposure–outcome association suffers from 'confounding'. The standard way to deal with confounding is to 'adjust for' (i.e. condition on) a selected set of potential confounders, for example by stratification or regression modeling. The rationale for this approach is that after adjustment, it may be reasonable to consider the exposure as being randomized 'by nature'. Formally, the aim of confounding adjustment is to produce conditional exchangeability:

$$\{Y_x\}_{x\in\mathbb{X}} \amalg X|C \tag{2.5}$$

Under consistency (2.1) and conditional exchangeability (2.5), $\Pr(Y = y|X = x, C) = \Pr(Y_x = y|C)$. It follows that any measure of the conditional association between X and Y, given C, equals the corresponding conditional population causal effect. For instance, $\Pr(Y = 1|X = 1, C)/\Pr(Y = 1|X = 0, C)$ equals $\Pr(Y_1 = 1|C)/\Pr(Y_0 = 1|C)$. The population (i.e. not C-specific) causal effect can be computed through 'standardization', i.e. by averaging over the marginal confounder distribution:

$$\Pr(Y_x = y) = E\{\Pr(Y = y|X = x, C)\}$$

2.4 Discussion

The potential outcome framework provides a natural definition of causal effects, based on the idea of comparing the same subjects under different exposure levels. Furthermore, it provides an elegant characterization of (conditional) exchangeability, which is the necessary requirement for drawing causal conclusions from observational data. During the last decades, the potential outcome framework has proven to be extremely useful, both in conceptualizing causal concepts (e.g. confounding and direct effects) and in the development of new methodology (e.g. propensity scores (Rosenbaum and Rubin, 1983), g-estimation (Robins, 1986), inverse probability weighting (Robins, 1997), and principal stratification (Frangakis and Rubin, 2002)).

One weakness of the potential outcome framework is that it is not so intuitive in the formulation and judgement of prior assumptions. Typically, subject matter experts understand well the causal structure of the problem (e.g. which the important confounders are and how the confounders are causally related to each other). In the potential outcome framework, however,

such expert knowledge must be recoded as independence statements involving counterfactual variables, as in (2.5). This task may be very difficult, since counterfactual independencies are rather technical conditions, which are often awkward to interpret. Pearl (2010) argued that in the potential outcome framework 'it is hard to ascertain whether all relevant judgments have been articulated, whether the judgments articulated are redundant, or whether those judgments are self-consistent'. Major progress was made when Pearl (1993) demonstrated that potential outcomes can be emulated through nonparametric structural equation (NPSE) systems. In an NPSE system, subject matter knowledge about underlying causal structures can be modeled explicitly, without having to be recoded into counterfactual independencies. All relevant counterfactual independencies can then be derived from the postulated causal model through simple algorithms. Pearl (2009) provides an exhaustive review of NPSEs and their relation to potential outcomes (see also Chapter 3 in this volume).

One common criticism of counterfactuals is that their empirical content is often somewhat vague. For example, suppose that we carry out a study to evaluate the effect of obesity on the risk for cardiovascular disease (CVD). A commonly used proxy for obesity is body mass index (BMI). Using this proxy we may, for example, formulate the research question as 'What is the effect of an increase in BMI from 25 to 35, on the risk for CVD?' In the potential outcome framework, this question calls for a comparison between two hypothetical scenarios: one where everybody has BMI equal to 25 and one where everybody has BMI equal to 35. But what does this mean? For any given subject, we can easily imagine a wide range of body compositions, all yielding a BMI equal to 35, but crucially different in other outcome-related aspects. A subject with BMI equal to 35 may, for instance, have either an extreme excess of body fat or an extreme excess of muscles. Since the research question allows for these crucially different interpretations, we may consider the question as being ill-posed and the underlying counterfactuals as being vague. The example highlights an important difference between association and causation. We would typically consider the BMI–CVD association to be a well-defined concept, because we agree on what it means to factually have a certain BMI level. However, in order for the causal effect to be well defined, this is not enough; we must also agree on what it means to have a *counter*factual BMI level, e.g. to have BMI equal to 35 when the factual BMI is equal to 25. This latter requirement is much stronger, and typically a source vagueness in causal effects, which is not present in measures of pure associations. One way to reduce such vagueness is to make precise the definition of obesity, including, for instance, the ratio of muscles to body fat, placement of fat on the body, etc. Under a sufficiently detailed definition of obesity, the research question may no longer be interpretable in crucially different ways.

Another option, which has been promoted by several authors (e.g. Robins and Greenland, 2000, and Hernán, 2005), is to make precise a hypothetical intervention able to produce the counterfactual scenarios of interest. Possible interventions that could potentially yield a specific BMI level are, for example, a strict diet or training program, surgery, genetic modification, etc. It is important to recognize, though, that although the latter strategy may result in a well-defined research question, it combines any 'intrinsic' effect of obesity with potential side-effects of the intervention. For instance, a training program may reduce a subject's risk for CVD, even if it fails to reduce the subject's BMI levels. Whether this feature is desirable or not clearly depends on the ultimate aim of the study. In practical scenarios, training programs may be the most realistic way to modify peoples BMI levels. Thus, if the aim is to guide policy makers rather than learning about the 'ethiology of obesity', then the effect of a training program may be a more suitable target for analysis than the effect of obesity per se. Chapter 9 in this volume provides a discussion of counterfactuals for nonmanipulable exposures.

In most applications, counterfactuals are assumed to be deterministic quantities. That is, if we could somehow 'play back time' and observe the same subject repeatedly under the same exposure level, it is assumed that we would always observe the same outcome. Whether this assumption can ever be reasonable has been debated in the causal inference literature. Pearl (2000) argued that the assumption follows naturally if we view a 'subject' as 'the sum of all experimental conditions that might possibly affect the individual's reaction, including biological, physiological, and spiritual factors, operating both before and after the application of the treatment'. Dawid (2000) argued that '... any attempt to refine the description of [subject] u still further, so as to recover the desired [deterministic] structure, surely risks becoming utterly ridiculous, and of no relevance to the kind of causal questions we really wish to address'. Although this debate may appear somewhat esoteric it has important practical implications; any results derived from a deterministic counterfactual model may be meaningless if the world is truly stochastic. Fortunately, some of the results that have been derived under deterministic counterfactual models have also been reproduced under models utlilizing stochastic couterfactuals (e.g. Dawid, 2003).

References

Dawid, A.P. (2000) Causal inference without counterfactuals. *Journal of the American Statistical Association*, **95** (450), 407–448.

Dawid, A.P. (2003) Causal inference using influence diagrams: the problem of partial compliance (with discussion), in *Highly Structured Stochastic Systems* (eds P.J. Green, N.L. Hjort and S. Richardson). Oxford University Press, pp. 45–81.

de Finetti, B. (1974) *Theory of Probability*. New York: John Wiley & Sons, Inc.

Frangakis, C.E. and Rubin, D.B. (2002) Principal stratification in causal inference. *Biometrics*, **58**, 21–29.

Galles, D. and Pearl, J. (1998) An axiomatic characterization of causal counterfactuals. *Foundations of Science*, **3** (1), 151–182.

Hernán, M.A. (2005) Invited commentary: hypothetical interventions to define causal effects–afterthought or prerequisite? *The American Journal of Epidemiology*, **162** (7), 618–620.

Hernán, M.A. and Robins, J.M. (2006) Instruments for causal inference: an epidemiologists dream? *Epidemiology*, **17** (4), 360–372.

Hernán, M.A. and Robins, J.M. (forthcoming) *Causal Inference*. Chapman & Hall/CRC.

Hudgens, M.G. and Halloran, M.E. (2008) Toward causal inference with interference. *Journal of the American Statistical Association*, **103** (482), 832–842.

Pearl, J. (1993) Aspects of graphical models connected to causality, in *Proceedings of the 49th Session of the International Statistical Institute*, Florence, Italy, Tome LV, Book 1, pp. 399–401.

Pearl, J. (2000) The logic of counterfactuals in causal inference (Discussion of 'Causal inference without counterfactuals' by A.P. Dawid). *Journal of the American Statistical Association*, **95** (450), 428–431.

Pearl, J. (2001) Direct and indirect effects, in *Proceedings of the Seventeenth Conference on Uncertainty in Artificial Intelligence*. (eds D. Koller and J. Breese). San Francisco, CA: Morgan Kaufmann, pp. 411–420.

Pearl, J. (2009) *Causality: Models, Reasoning and Inference*, 2nd edn. Cambridge: Cambridge University Press.

Pearl, J. (2010) An introduction to causal inference. *The International Journal of Biostatistics*. **6** (2), Article 7.

Robins, J.M. (1986) A new approach to causal inference in mortality studies with sustained exposure periods–application to control of the healthy worker survivor effect. *Mathematical Modelling*, **7**, 1393–1512.

Robins, J.M. (1989) The analysis of randomized and non–randomized AIDS treatment trials using a new approach to causal inference in longitudinal studies, in *Health Service Research Methodology: a focus on AIDS* (eds L. Sechrest, H. Freeman and A. Mulley) pp. 113–159.

Robins, J.M. (1997) Marginal structural models. In *Proceedings of the American Statistical Association. Section on Bayesian Statistical Science*, pp. 1–10.

Robins, J.M. and Greenland, S. (2000) Discussion of 'Causal inference without counterfactuals' by A.P. Dawid. *Journal of the American Statistical Association*, **95** (450), 431–435.

Rosenbaum, P.R. and Rubin, D.B. (1983) The central role of 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** (5), 688–701.

Rubin, D.B. (1978) Bayesian inference for causal effects: the role of randomization. *The Annals of Statistics*, **6** (1), 34–58.

Rubin, D.B. (1980) Discussion of 'Randomization analysis of experimental data: the Fisher randomization test' by D. Basu. *Journal of the American Statistical Association*, **75**, 591–593.

Tchetgen Tchetgen. E.J. and VanderWeele, T.J. (2010) On causal inference in the presence of interference. *Statistical Methods in Medical Research*. DOI: 10.1177/0962280210386779.

3

Structural equations, graphs and interventions

Ilya Shpitser
Department of Epidemiology, Harvard School of Public Health, Boston, Massachusetts, USA

3.1 Introduction

Causality is fundamental to our understanding of the natural world. Causal statements are a part of everyday speech, as well as legal, scientific and philosophical vocabulary. Human beings reach an intuitive consensus on the meaning of many causal utterances and there have been numerous attempts to formalize causality in a way that is faithful to this consensus (Wright, 1921; Neyman, 1923; Tinbergen, 1937; Lewis, 1973; Rubin, 1974; Robins, 1987; Pearl, 2000).

The notion central to most of these formalizations is that of an intervention – a kind of idealized randomized experiment imposed on a 'set of units', which can represent patients in a medical trial, a culture of cells, and so on. Many scientific questions can be phrased in terms of effects of such experiments. Two influential formalizations of causation based on interventions are the framework of potential outcomes (Neyman, 1923; Rubin, 1974) and the framework of nonparametric structural equation models (NPSEMs) (Pearl, 2000).

The potential outcome framework, discussed by Sjölander in Chapter 2 of this volume, considers the effect of an intervention on a *treatment variable A* on an *outcome variable Y*. The resulting potential outcome, which we – in accord with the previous chapter – shall denote by Y_a, or $Y(a)$, is a random variable representing the behavior of Y in a hypothetical situation where the value of a random variable A was set to a, without regard for the normal behavior of A, a setting that may be implemented by randomization in practice. In the potential outcome framework, causal questions are encoded as statements about joint distributions made up

Causality: Statistical Perspectives and Applications, First Edition. Edited by Carlo Berzuini, Philip Dawid and Luisa Bernardinelli.
 Published 2012 by John Wiley & Sons, Ltd.

of potential outcome random variables, and causal assumptions are expressed as various restrictions on these distributions.

Nonparametric structural equation models define a very general data-generating mechanism suitable for encoding causation. In such models observable variables are functions of other observable variables, and possibly of noise terms, with the entire model behaving as a kind of stochastic circuit. An intervention on a variable A in an NPSEM replaces the function determining the value of A by a constant function, resulting in a new circuit. Causal relationships between variables in NPSEMs can be summarized by means of a graph called a causal diagram, with some causal assumptions entailed by the model represented by missing edges in the graph. NPSEMs induce joint distributions over pre- or post-intervention variables. These distributions provide a close link between the NPSEM formalism and the potential outcome formalism.

In this chapter we introduce the NPSEMs framework along with the relevant background from graph theory and probability theory, show the close connection between structural equations and potential outcomes, and argue for the conceptual complementarity of the two frameworks, with NPSEMs providing transparency and visual intuition of graphs, and potential outcomes giving a coherent way of thinking about causation in the "epistemically impoverished" situations frequently encountered in scientific practice.

3.2 Structural equations, graphs, and interventions

3.2.1 Graph terminology

We first introduce graph-theoretic terminology which is necessary for discussing NPSEMs in this chapter, and identifiability problems in Chapter 6 of this volume. In mathematics and computer science, graphs are used to model objects, represented by nodes, and their relationships, represented by edges. In this chapter, we will restrict our attention to graphs where the edges are either directed (contain a single arrowhead), or bidirected (contain two arrowheads). More specifically, we shall restrict our attention to directed and mixed graphs. A *directed graph* consists of a set of nodes and directed arrows connecting pairs of nodes. A *mixed graph* consists of a set of nodes and directed and/or bidirected arrows connecting pairs of nodes. A *path* is a sequence of distinct nodes where any two adjacent nodes in the sequence are connected by an edge. A *directed path* from a node X to a node Y is a path where all arrows connecting nodes on the path have an arrowhead pointing away from X and towards Y. If an arrow has a single arrowhead pointing from X to Y, then X is called a *parent* of Y and Y a *child* of X. If X and Y are connected by a bidirected arrow, they are called *spouses*. The set of spouses of a node X is denoted by $Sp(X)$. If X has a directed path to Y then X is an *ancestor* of Y, and Y a *descendant* of X. Non-descendants, descendants, ancestors, parents, and children of X are denoted, respectively, by $Nd(X)$, $De(X)$, $An(X)$, $Pa(X)$, and $Ch(X)$. By convention, X is both an ancestor and a descendant of X. A directed acyclic graph (DAG) is a directed graph where for any directed path from X to Y, Y is not a parent of X. An *acyclic directed mixed graph* (ADMG) is a mixed graph, which is a DAG if restricted to directed edges.

A consecutive triple of nodes W_i, W_j, W_k on a path is called a *collider* if the edge from W_i to W_j and the edge from W_k to W_j both have arrowheads pointing to W_j. Any other consecutive triple is called a *noncollider*. A path between two nodes X and Y is said to be *blocked* by a set Z if either for some noncollider on the path, the middle node is in Z, or for some collider

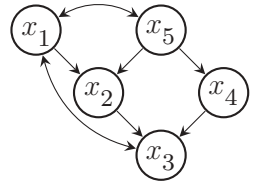

Figure 3.1 An acyclic directed mixed graph.

on the path, no descendant of the middle node is in Z. For disjoint sets X, Y, Z of nodes in an ADMG we say X is m-separated from Y given Z if every path from a node in X to a node in Y is blocked by Z. If the ADMG is also a DAG, then we say X is d-separated from Y given Z. If X is not m-separated (or d-separated) from Y given Z, we say X is m-connected (or d-connected) to Y given Z. See the graph in Figure 3.1 for an illustration of these concepts. In this graph:

- $X_1 \rightarrow X_2 \rightarrow X_3 \leftarrow X_4$ is a path from X_1 to X_4;
- $X_1 \rightarrow X_2 \rightarrow X_3$ is a directed path from X_1 to X_3;
- X_1 is a parent of X_2 and an ancestor of X_3;
- $X_2 \rightarrow X_3 \leftarrow X_4$ is a collider;
- X_1 is m-separated from X_4 given X_5;
- X_1 is m-connected to X_4 given X_5 and X_3.

3.2.2 Markovian models

A Markovian model, also known as a nonparametric structural equation model (NPSEM), is a tuple $\langle U, V, F, P(u)\rangle$, where the tuple elements are defined as follows. The symbol V denotes a set of observable random variables and the symbol U denotes a set of unobserved background random variables with one $U_i \in U$ assigned to each $V_i \in V$. The symbol F denotes a set of functions, one for each element V_i in V. Each $f_i \in F$ corresponding to $V_i \in V$ maps values of a subset of $V \setminus \{V_i\}$ and of the background variable $U_i \in U$ corresponding to V_i to values of V_i. Finally, $P(u) = \prod_i P(u_i)$ is a joint distribution over U. This distribution $P(u)$, along with F, induces a joint distribution $P(v)$ over V. A Markovian model may be represented by a directed graph, where there is a vertex for every random variable in V, and for two random variables $V_i, V_j \in V$, there is a directed edge $V_i \rightarrow V_j$ if and only if the subset of V whose values are mapped on to values of V_j by f_j contains V_i. This graph is said to be induced by the Markovian model and is called the *causal diagram* representation of the NPSEM. When the causal diagram is acyclic, the corresponding Markovian model is called *recursive*. In the remainder of this section, we will restrict our attention to recursive models.

As is the case in a standard Bayesian network (Pearl, 1988), the distribution $P(v)$ under a Markovian model is Markov relative to its induced graph $\mathcal{G}$. In other words, we have $P(v) = \prod_i P(v_i|pa(v_i)_{\mathcal{G}})$. This is equivalent to the local Markov property, which states that each V_i is independent of $V \setminus (De(V_i)_{\mathcal{G}} \cup Pa(V_i)_{\mathcal{G}})$ given $Pa(V_i)_{\mathcal{G}}$, and in turn equivalent to the global Markov property defined by d-separation (Pearl, 1988), which states that if X is

d-separated by Y given Z in the model-induced graph, then X is independent of Y given Z in the joint distribution $P(v)$ of the model variables. The importance of directed graphs in both Bayesian networks and Markovian models stems from this relationship between d-separation, a path property defined on a graph, and probabilistic independence in an appropriate joint distribution.

Unlike Bayesian networks, Markovian models permit formalization of causality by means of the *intervention operation*, denoted by $do(v_i)$ in Pearl (2000), and defined as the action of setting the value of V_i to be equal to v_i by an external intervention. This operation, applied to a Markovian model M, constructs a *submodel* M_{v_i}, which is obtained from M by replacing the function f_i determining V_i in M by a new constant function f_i^* that always gives value v_i, more formally defined as $(\forall x) f_i^*(x) = v_i$. Interventions represent idealized experiments. The result of such an experiment, called the *causal effect* of $do(v_i)$, is the distribution induced by the intervention on the remaining variables $V \setminus \{V_i\}$ in M_{v_i}, this distribution often being called 'interventional', and denoted as $P(v \setminus \{v_i\}|do(v_i))$ or $P_{v_i}(v \setminus \{v_i\})$. The causal effect of $do(v_i)$ on a subset of interest $Y \subset V \setminus \{V_i\}$ is denoted by $P(y|do(v_i))$ or $P_{v_i}(y)$. Note that, in the causal inference literature, the so-called population causal effect of interest may actually be a contrast between two intervention distributions, induced by setting V_i at v_i and v_j, respectively. Examples of contrasts are the causal mean difference $E(y|do(v_i)) - E(y|do(v_j))$ or the causal relative risk $P(y|do(v_i))/P(y|do(v_j))$. However, these quantities can all be computed from $P(y|do(v_i))$ if this distribution is known for all exposure levels v_i. Thus, without loss of generality, we will refer to $P(y|do(v_i))$ as the causal effect of the intervention $do(v_i)$.

The causal diagram induced by M_{v_i} can be obtained from the causal diagram G induced by M by removing all arrows pointing to V_i in G. We will denote the resulting graph, called the *mutilated graph* in Pearl (2000), by $G_{\overline{v_i}}$. Since the variable V_i after the intervention $do(v_i)$ is a constant, it is sometimes omitted from the mutilated graph, in which case the graph representing M_{v_i} is simply obtained by removing V_i and all edges adjacent to V_i from G.

Many scientific questions can be formalized as being about causal effects in an appropriate Markovian model. In practice, interventions can be implemented by means of an actual randomized experiment, which allows the intervention distributions of interest to be estimated directly from the generated experimental data. The difficulty is that experimentation is often expensive, impractical, or illegal, which leads naturally to the question of whether causal effects can be expressed in terms of the observational distribution in a Markovian model, since this distribution can be estimated just by passive observation of a set of collected samples, rather than by performing experiments. This is a question of *identification* of causal effects from $P(v)$. In Markovian models, the effect $P(v \setminus \{v_i\}|do(v_i))$ is always identifiable and equal to

$$P(v \setminus \{v_i\}|do(v_i)) = \frac{P(v)}{P(v_i|pa(v_i))}$$

where it is assumed the value v_i of V_i is consistent with values v of V. This is known as the truncation formula (Spirtes *et al.*, 1993; Pearl, 2000), or the g-formula (Robins, 1986). This formula generalizes in an obvious way to interventions on multiple variables and subsets of $V \setminus \{V_i\}$. In particular, for any V_j, $P(v_j|do(pa(v_j))) = P(v_j|pa(v_j))$.

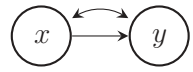

Figure 3.2 A graph representing semi-Markovian models where $P(y|do(x))$ is not identifiable from $P(v)$.

3.2.3 Latent projections and semi-Markovian models

Latent variables are very common in causal inference problems. For this reason it is useful to consider Markovian models where a subset of variables in V are unobserved. Such models can be represented by a directed acyclic graph with a subset of nodes labeled as latent. However, another alternative is to represent such models by acyclic directed mixed graphs (ADMGs) containing both directed arrows, representing "direct causation", and bidirected arrows, representing 'unspecified causally ancestral latent variables'. For example, in the ADMG shown in Figure 3.2, the X variable is a direct cause of Y and, in addition, these two variables share unspecified common causes. Every Markovian model with a set of nodes marked latent can be represented by an ADMG called a *latent projection*.

Definition 1. Latent projection *Let M be a Markovian model inducing $\mathcal{G}$, where V is partitioned into a set L of latents and a set O of observables.*

Then the latent projection $\mathcal{G}(O)$ is a graph containing a vertex for every element in O, where for every two nodes $O_i, O_j \in O$:

- *$\mathcal{G}(O)$ contains an edge $O_i \rightarrow O_j$ if there exists a d-connected path $O_i \rightarrow I_1 \rightarrow \cdots \rightarrow I_k \rightarrow O_j$, where $I_1,...I_k \in L$.*
- *$\mathcal{G}(O)$ contains an edge $O_i \leftrightarrow O_j$ if there exists a d-connected path $O_i \leftarrow I_1...I_k \rightarrow O_j$, $O_i \leftarrow I_1...I_k \leftrightarrow O_j$, or $O_i \leftrightarrow I_1,..., I_k \rightarrow O_j$, where $I_1,..., I_k \in L$.*

Latent projections are a convenient way to represent latent variable Markovian models because they simplify the representation of multiple paths of latents while preserving many useful constraints over observable variables.

An alternative way to think about latent projections is to consider a generalization of Markovian models that induces ADMGs directly. This generalization is the semi-Markovian model (Pearl, 2000). Like Markovian models, semi-Markovian models are tuples $\langle U, V, F, P(u)\rangle$. The sole difference is we allow the values of a single $U_k \in U$ to influence either one or two functions in F, unlike the Markovian case where values of a single U_i influenced a single function $f_i \in F$. The induced graph of a semi-Markovian model contains a vertex for every element in V, with two elements V_i, V_j connected by a directed arrow as before, and connected by a bidirected arrow if f_i, f_j are both influenced by values of a single $U_k \in U$.

3.2.4 Interventions in semi-Markovian models

Since semi-Markovian models are causal, the definitions of intervention and causal effect carry over from Markovian models without change. However, the identification problem of causal effects in semi-Markovian models is significantly more difficult than in Markovian models.

In particular, there exist ADMGs such that in some semi-Markovian models inducing these ADMGs certain causal effects are not identifiable. The simplest such graph is known as the "bow arc graph" and is shown in Figure 3.2. It is possible to construct two models M_1 and M_2 that will both induce the shown graph and have the same distributions $P(v)$, but which will disagree on the value of the causal effect $P(y|do(x))$. Hence the causal effect is not a function of the observed data in the class of model inducing the graph shown in Figure 3.2, and thus cannot be estimated without additional assumptions.

The natural question is the characterization of ADMGs where a particular causal effect is identifiable. This problem has received much attention in literature, with complete solutions given in Tian and Pearl (2002), Huang and Valtorta (2006), Shpitser and Pearl (2006). Chapter 6 in this volume will review the identification problem of causal effects and its solutions in greater detail.

3.2.5 Counterfactual distributions in NPSEMs

Markovian and semi-Markovian models are very general data generating mechanisms that can be viewed as stochastic circuits, where nodes in the model represent input, output, and intermediate wires, and functions that determine the value of variables are logic gates. In such models, the joint distribution representing the effect of $do(x)$ on a singleton outcome variable Y, represented using the do notation as $P(y|do(x))$, can also be represented in the language of potential outcomes as a random variable Y_x or $Y(x)$. We will call such variables *counterfactual*.

Note that if we fix all U variables in a semi-Markovian model, the values of observable variables in V become deterministically fixed. This holds true even after interventions. In other words, if we fix the unobservable variable set U to have values u, then the counterfactual random variable Y_x becomes a constant denoted by $Y_x(u)$.[1]

This allows us to use the distribution $P(u)$ to define joint distributions over counterfactual variables, even if the interventions that determine these variables disagree with each other. The precise definition is as follows:

$$P(Y^1_{x^1} = y^1, ..., Y^k_{x^k} = y^k) = \sum_{\{u|Y^1_{x^1}(u)=y^1 \wedge ... \wedge Y^k_{x^k}(u)=y^k\}} P(u) \tag{3.1}$$

where U is the set of unobservable variables in the model. In words, this definition says that a joint probability of counterfactual variables $Y^1_{x^1}, ..., Y^k_{x^k}$ assuming values $y^1, ..., y^k$ is defined to be the sum of probabilities (according to $P(u)$) of values u of unobservable variables U, which result in constants $Y^1_{x^1}(u), ..., Y^k_{x^k}(u)$ being equal to $y^1, ..., y^k$.

The value of such distributions is that they allow us to formalize counterfactual reasoning. Consider the following question: "I have a headache. Would I have a headache had I taken an aspirin one hour ago?" Questions of this type are very natural[2] and frequently arise both in informal and scientific discourse. In Daurd (2007) they are called "questions about the causes of effects". We can formalize this question by considering a Markovian model M with two variables: A (representing the decision to take aspirin one hour ago) and Y (for headache now).

[1] Interpreting value assignments u as 'units', this implies that the stable unit treatment value assumption (SUTVA) holds in NPSEMs.

[2] One piece of evidence for the naturalness of such questions, at least in English, is that the English counterfactual connective 'had' is a short word.

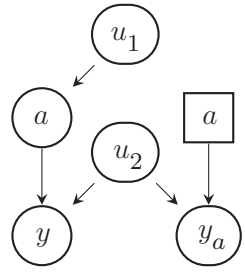

Figure 3.3 A counterfactual graph for the aspirin example. The square node represents an intervened-on node set to a constant.

Both variables are binary, with value 1 representing aspirin taken and headache present, and value 0 representing aspirin not taken and headache absent. It is reasonable to assume that the decision to take aspirin has some sort of effect on headache, but not vice versa, so our causal diagram will have A as the parent of Y. The question we are trying to formalize is referencing two possible worlds, the 'actual world,' where headache happened, and the hypothetical world, which is 'as close as possible' to the actual world, except for the fact that aspirin was taken one hour ago. The hypothetical world is represented by a submodel $M_{a=1}$. The expression "headache had I taken aspirin" refers to the variable Y in this submodel, in other words to $Y_{a=1}$. The original question also had a statement 'I have a headache,' referencing the variable Y in the actual world, represented by the model M.

Hypothetical worlds being 'as close as possible' to the actual world is modelled in NPSEMs by having all counterfactual worlds share unobservable variables. These variables represent the causal past, which lead to the currently observable state of affairs, with interventions representing minimal relevant differences between worlds. There is a graphical representation of these connected worlds known as the *counterfactual graph* (Shpitser and Pearl, 2007). Such a graph contains copies of the original causal diagram, one copy for each counterfactual world of interest. A copy corresponding to a hypothetical world where the intervention $do(x)$ took place is actually the mutilated graph $G_{\overline{x}}$. Each such copy shares the unobservable variables with other copies. In our aspirin example, the counterfactual graph is shown in Figure 3.3. See Shpitser and Pearl (2007) for technical details of constructing counterfactual graphs for arbitrary counterfactual queries. The counterfactual graph for the special case of two hypothetical worlds, as in our example, has been proposed in Balke and Pearl (1994a, 1994b) and is referred to as the *twin network graph*.

We can view the counterfactual graph in our example as a causal diagram representing another NPSEM, which contains both counterfactual worlds. In particular, d-separation in the counterfactual graph implies conditional independences in a joint distribution over counterfactual (potential outcome) variables $P(A, Y, Y_{a=1})$, defined as in Equation (3.1). Our counterfactual question: 'I have a headache. Would I have had a headache had I taken an aspirin one hour ago?' can now be represented as a conditional distribution derived from the above joint, specifically by the probability $P(Y_{a=1}|Y = 1)$. Identification of such distributions from either interventional or observational data, assuming a complete causal diagram is known, is sometimes possible. A treatment of this version of the identification problem appears in the literature (Shpitser and Pearl, 2007).

3.2.6 Causal diagrams and counterfactual independence

One of the advantages of using NPSEMs to talk about potential outcomes is that a fully specified causal diagram is an oracle that makes causal assumptions explicit as missing edges.

Consider an ADMG G representing a semi-Markovian model. For any node X in G, it is the case that $X_{pa(x)} = X_{nd(x)}$. In words, this assumption states, for all X, that any direct arrow missing from a nondescendant W of X to X implies fixing W to have no effect on X as long as all parents of X are already fixed. These assumptions are called exclusion restrictions in Pearl (2000). Similarly, for any missing bidirected arrow from W to X, it is the case that $W_{pa(w)}$ is independent of $X_{pa(x)}$. In words, two counterfactual variables (with their respective parents fixed to any value) cannot be dependent if the original causal diagram does not connect them by a bidirected arc. These assumptions are called independence restrictions in Pearl (2000).

These assumptions can be used to derive other counterfactual independences by means of an important axiom, which we call generalized consistency, but which is called composition in Pearl (2000). This axiom, which is generally assumed to hold in the potential outcome framework, is also assumed to hold in NPSEMs. Formally, it states that

$$(\forall u) W_x(u) = w \Rightarrow (Y_x(u) = Y_{x,w}(u))$$

In words, for any value assignment u to unobservable variables U, if the counterfactual variable $W_x(u)$ assumes value w, then the values assumed by counterfactual variables $Y_x(u)$ and $Y_{x,w}(u)$ will be the same. This assumption can be viewed as restating the convention of the sharing of unobserved variables between hypothetical worlds making up the counterfactual graph.

Coupled with generalized composition, exclusion and independence restrictions can be used logically to derive all independences in any joint distribution over counterfactual variables obtained from an NPSEM (Halpern, 2000). In some sense, these restriction form a kind of "logical basis" for counterfactual constraints implied by an NPSEM. Causal diagrams not only make this basis explicit via missing edges, but can also vastly simplify derivation of implications of this basis. Indeed, the counterfactual graphs of the kind shown in the last section can be used to show counterfactual independence directly by means of a graphical d-separation test. For example, we can conclude using d-separation that Y_a is independent of A, but dependent of A given Y. In complex causal models, this sort of counterfactual independence test is very difficult to perform without graphical aid.

3.2.7 Relation to potential outcomes

An NPSEM with a given causal diagram G implicitly defines counterfactual distributions over variables that can be viewed as potential outcomes. The standard assumptions made in the potential outcome framework also hold in the NPSEM framework. In particular, the SUTV assumption and the generalized consistency (composition) assumption hold in NPSEMs.

The advantage of NPSEMs is making assumptions explicit by means of graphs. Graphical language is more easily understood by the human mind, and counterfactual independence tests are much easier to verify via graphs than via unaided causal judgement. Furthermore, NPSEMs, as stochastic circuits, give a very reasonable data generating mechanism for a wide range of processes encountered in science. On the other hand, the reality of the scientific process is that unsettled areas of empirical science, by their very nature, suffer from lack of reliable

knowledge of causality. This lack of knowledge is precisely what is motivating scientific investigation in the first place. When causal knowledge is scarce, it is not generally possible to justify causal assumptions needed to construct the full graph necessary for reasoning with NPSEMs. Instead, scientists may be justified in making a handful of assumptions on the data generating process, such as conditional ignorability (Rosenbaum and Rubin, 1983), which may be just enough to identify one causal quantity of interest, and little else.

One effective compromise between the "epistemically extravagant" NPSEMs and minimal assumptions made in the potential outcomes literature are the so-called minimal causal models (MCMs) (Robins and Richardson, 2011). These models are similar in spirit to NPSEMs in the sense of representing independences by means of graphs. At the same time, MCMs never make counterfactual independence claims that cannot be tested in principle, which is the cautious stance reminiscent of the practitioner willing to adopt potential outcomes, but not willing to commit to a detailed causal theory.

An alternative compromise is to use the detailed theory of identification developed in NPSEMs as a springboard to further results that relax graphical assumptions as much as possible while still yielding identification of causal quantities of interest. One recent result in this vein characterized assumptions necessary for covariate adjustment for identifying causal effects (Shpitser *et al*., 2010).

References

Balke, A. and P. J. (1994a) Counterfactual probabilities: computational methods, bounds and applications, in *Proceedings of UAI-94*, pp. 46–54.

Balke, A. and Pearl, J. (1994b). Probabilistic evaluation of counterfactual queries, in *Proceedings of AAAI-94*, pp. 230–237.

Dawid, A. P. (2007) Countertactuals, hypotheticals and potential responses: a philosophical examination of statistical causality, in *Causality and Probability in the Sciences*, Texts in Philosophy, vol. 5 (eds F. Russo and J. Williamson). London: College Publications, pp. 503–532.

Halpern, J. (2000) Axiomatizing causal reasoning. *Journal of A.I. Research*, pp. 317–337.

Huang, Y. and Valtorta, M. (2006) Identifiability in causal Bayesian networks: a sound and complete algorithm, in *Twenty-First National Conference on Artificial Intelligence*, 2006.

Lewis, D. (1973) *Counterfactuals*. Cambridge, MA: Harvard University Press.

Neyman, J. (1923) Sur les applications de la thar des probabilities aux experiences agaricales: essay des principle. Excerpts reprinted (1990) in English. *Statistical Science*, **5**, 463–472.

Pearl, J. (1988) *Probabilistic Reasoning in Intelligent Systems*. San Mateo, CA: Morgan and Kaufmann.

Pearl, J. (2000) *Causality: Models, Reasoning, and Inference*. Cambridge: Cambridge University Press.

Robins, J.M. (1987) A graphical approach to the identification and estimation of causal parameters in mortality studies with sustained exposure periods. *Journal of Chronic Disease*, **2** 139–161.

Robins, J.M. and Richardson, T.S. (2011) Alternative graphical causal models and the identification of direct effects, in *Causality and Psychopathology: finding the determinants of disorders and their cures*.

Robins, J.M. (1986) A new approach to causal inference in mortality studies with sustained exposure periods – application to control of the healthy worker survivor effect. *Mathematical Modeling*, **7**, 1393–1512.

Rosenbaum, P.R. and Rubin, D.B. (1983) The central role of the propensity score in observational studies for causal effects. *Biometrika*, **70** (1), 41–55.

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

Shpitser, I. and Pearl, J. (2006) Identification of conditional interventional distributions. in *Uncertainty in Artificial Intelligence*, vol. 22.

Shpitser, I. and Pearl, J. (2007) Complete identification methods for the causal hierarchy. Technical Report R–336, UCLA Cognitive Systems Laboratory.

Shpitser, I. VanderWeele, T. and Robins, J.M. (2010) On the validity of covariate adjustment for estimating causal effects, in *International Joint Conference on Artificial Intelligence*, vol. 22.

Spirtes, P. Glymour, C. and Scheines, R. (1993) *Causation, Prediction, and Search*. New york: Springer Verlag.

Tian, J. and Pearl, J. (2002) A general identification condition for causal effects, in *Eighteenth National Conference on Artificial Intelligence*, pp. 567–573.

Tinbergen, J. (1937) *An Econometric Approach to Business Cycle Problems*. Paris: Hermann.

Wright, S. (1921) Correlation and causation. *Journal of Agricultural Research*, **20**, 557–585.