REFERENCE ONLY

UNIVERSITY OF LONDON THESIS

Degree PhD Year 2005 Name of Author GENELETIS G.

This thesis comes within category D.

☑ This copy has been deposited in the Library of UCL

☐ This copy has been deposited in the Senate House Library, Senate House, Malet Street, London WC1E 7HU.

Aspects of Causal Inference in a non-counterfactual Framework

Thesis submitted to the University of London for the degree of Doctor of Philosophy in the Faculty of Science

by

Sara Gisella Geneletti

Department of Statistical Science
University College London

August 2005

ProQuest LLC
789 East Eisenhower Parkway
P.O. Box 1346
Ann Arbor, MI 48106-1346

First and foremost, I dedicate this Thesis to Wilma and Carlo for 29 years of moral and financial support. Grazie

Acknowledgements

I would like to acknowledge my supervisors Phil Dawid and Vanessa Didelez for the inspiration and support they have been over the last long three and some years.

I'd like thank my fellow Phd students, faculty and staff in the Department of Statistical Science who have been of great help.

Also my friends, Charlotte, Katia, Stew, Enrica, Andrew in London and Corina, Hibba, Marielle, Laura, Alina, Emi who are scattered around the world. Last but not least *thanks D!*

Abstract

Since the mid 1970s and increasingly over the last decade, causal inference has generated interest and controversy in statistics. Mathematical frameworks have been developed to make causal inference in fields ranging from epidemiology to social science. However, most frameworks rely on the existence of counterfactuals. and the assumptions that underpin them are not always made explicit. This thesis analyses such assumptions and proposes an alternative model. This is then used to tackle problems that have been formulated in counterfactual terms.

The proposed framework is based on decision theory. Causes are seen in terms of interventions which in turn are seen as decisions. Decisions are thus explicitly included as *intervention variables*, in both algebraic expressions for causal effects and the in DAGs which represent the probabilistic structure between the variables.

The non-counterfactual framework introduces a novel way of determining whether causal quantities are identifiable. Two such quantities are considered and conditions for their identification are presented. These are the *direct effect* of treatment on response in the presence of a mediating variable, and the *effect of treatment on the treated.* To determine whether these are identifiable, intervention nodes are introduced on the variables that are thought to be causal in the problem. By manipulating the conditional independences between the observed variables and the intervention nodes it is possible to determine whether the quantities of interest can be expressed in terms of the

a) specific settings and/or *b)* the idle setting of the intervention nodes, corresponding to experimental regimes and the observational regimes of the causal variables.

This method can be easily tailored to any specific context, as it relies only on the understanding of conditional independences.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

The original work in this thesis is based on the decision theoretic framework described in Lindley (1985) and Raiffa (1970) and introduced as a method for causal inference in Dawid (2000).

The objectives are to formally develop the decision theoretic framework for causal inference, to assess the validity of assumptions underlying the frameworks in the literature, and finally, to develop some aspects of causal inference using the decision theoretic framework.

Before launching into the thesis, a few points worth noting. The first is that there is no data analysis in this thesis. It deals purely with building the methodology and applying it to specific problems at a pre-data analytic stage.

The second point is a brief comment on why I am interested in causal inference; It is the fundamental aim of scientific research and thus worthy of pursuit. Also, causal inference is a grey area between what is considered "hard science" and metaphysics.

After reading the basic literature in the field, I realised that this aspect of causal inference is played down if it is mentioned at all. This seemed to be somewhat of a paradox. Causal inference had been neglected in statistics for a very long time, precisely because of its elusive qualities, why, now that it had been resurrected, was no one referring to them? Perhaps this was in order to give causal inference the hard science reputation it had lacked so far. However, all inference is based on assumptions that cannot be tested, maybe what we consider hard science is just what is based on the most accepted assumptions. Whatever the reasons, in this thesis I have space and time to dedicate to the grey area in causal inference.

Finally, the arguments covered in this work explore different facets of causal inference, and are connected by the following four themes: *(a)* the assumptions underlying causal models and their applications, *(b)* translating methods initially formulated in counterfactual frameworks in terms of the decision theoretic framework developed in Chapter 2, *(c)* the relationship between data collected under passive observational regimes and under experiment, and finally, *(d)* identifiability issues. Clearly these themes are linked to one another and in particular the last two go hand in hand.

Chapter 2 has two parts. The first part is an informal metaphysical discussion that motivates the choice of the decision theoretic framework for causal inference. We discuss what we believe cause to be and the role of intervention in our understanding of causality. The second part is a formal description of the decision theoretic framework, and notation, algebra and graphical con-

cepts are introduced.

In Chapter 3, the competing causal inference frameworks are briefly described and discussed. We start with Rubin's potential outcomes framework, then cover functional model-graphical frameworks, focusing in particular on Pearl's causal model and Heckerman and Shachter's decision theoretic causal model, not to be confused with the framework developed in Chapter 1, as it is not truly decision theoretic. Finally, the frameworks are compared and the assumptions they are based on are questioned.

Chapter 4 looks at the assumptions on which causal discovery algorithms are based. These algorithms mine large databases of observational (passively gathered) data for causal relationships. In particular, we focus on the implicitly made assumption that it is possible discover causal relationships from data gathered in observational studies, which we term the *fundamental assumption.*

When the response to a particular treatment is thought to be partially mediated by another factor, it is often of interest to determine how much is mediated, that is the *indirect* effect and how much is instead purely due to the treatment, that is the *direct effect.* In Chapter 5 we develop the problem of identifying direct and indirect effects of treatments in non-counterfactual terms. Of particular interest is the development of the *manipulation* variable which is a fictional variable (it need not be feasible in reality) that enables the identification of direct and indirect effects from purely observational data.

In Chapter 6 the problem of the *effect of treatment on the treated* (ETT)

is tackled in the decision theoretic framework. This problem arises when the data are *(i)* observational studies where only the response and covariates of the treated are available or *(ii)* experimental data without a control group, and it is thus not possible to estimate the average treatment effect which is normally the causal quantity of interest. The ETT depends on an unobservable selection criterion. We look in detail at whether the ETT is well-defined for different values of the unobservable criterion, and also at how to identify the ETT.

Finally, in Chapter 7, we summarise the results and discuss further research.

The first part of Chapter 2, up to section 2.3, as well as Chapters 4,5 and 6 are original work developed by me with the aid of my supervisor. Chapter 3 is principally a description of competing causal frameworks. Translations into the Decision theoretic framework however, are again my work.

Chapter 2

The Decision Theoretic Causal Model

2.1 Causal Concepts

In order to give an informal basis to the framework of causal inference proposed in this thesis, I will begin by considering in simple terms how people express causal concepts and what they believe them to be. These discussions are of a philosophical nature and can be treated as separate from the later formal development of the causal inference methodology, which starts from section 2.2 onwards. However, I believe that this is a necessary exercise when studying this subject, as divergences in this area lead to different specifications of methodology. How my beliefs about the nature of causality are in part responsible for my choice of methodology is discussed in section 2.1.3.

2.1.1 How we express cause

Causal concepts are expressed in many different ways both in colloquial language and in more formal contexts. Below is a list of different expressions, some involving the word *cause* directly and others implying causal relationships. Most are based on Dawid (2000) and ensuing comments.

2.1.1 Causal expressions

1) My headache is gone because I took an aspirin.

2) If I had taken an aspirin my headache would have gone away.

3) Will my headache go away if I take an aspirin?

4) The central question in any employment discrimination case is whether the employer would have taken the same action had the employee been of a different race (age, sex, religion, national origin etc.) and everything else had been the same. (Carson versus Bethlehem Steeo Corp., 70 EEP Cases 921,7th Cir.(1996), Quoted in Gaswirth 1997. (Pearl 2001b)

5) She did not get the promotion because she is a woman.

6) The pavement is wet because it rained.

It is worth noting a few points. Expression (2) is a *counterfactual* i.e., it asks what would have happened if counter to fact, the person had taken aspirin. It is impossible to verify what the consequences of taking aspirin would have

been and hence to verify the truth of such an expression. Expression (1) although not explicitly counterfactual, operates on a similar principle. It implies that if the person had not taken an aspirin then the headache would not have gone away. That is, it compares what did happen to (the unknowable) what would have happened had a different action been taken. This concept applies to expressions (5) and (6) as well, however, these are fundamentally different in that they attribute cause to natural phenomena, the sex of the person who did not get promoted and the rain, not the voluntary action of taking or not an aspirin. Expression (4) focuses on comparing what would have happened if a person had been of a different sex, race etc, given that all else is the same. However, it does not go so far as to attributing cause directly to the sex, race etc. It is in fact the defendant who is potentially responsible for the discrimination and has to ultimately pay the fine if found guilty. Finally consider expression (3). It is different from the rest in that it is not a statement, but a question. It seeks to verify whether the action of taking aspirin performed now is a cause of change in headache status in the future.

Prospective vs Retrospective causality

Another important distinction to be made in causal expressions it that between *retrospective* and *predictive* questions. We can either ask whether the aspirin been taken already, that is, *Sara has taken an aspirin, and her headache is gone, is it because she took the aspirin?* Or the aspirin has not

been taken and the question is *Sara has a headache, will taking the aspirin make the headache go away?* Although they are similar, these questions address different issues. The first question is retrospective, it asks if it can attribute a present non-intentional change in status (that is Sara's headache) to a decision made sometime in the past. The second is predictive, it asks whether, with respect to other possible decisions (in this case not taking the aspirin), the decision to take an aspirin will result in the desired effect, that is, the headache going away.

2.1.2 How we perceive cause

Given how we express causal concepts, what qualifies a relationship as causal? This question has two types of answers. The first concerns the mental process that takes place when human beings elaborate what they *perceive* or *sense* (see, feel, hear etc) in the world around them. This process results in some relationships being judged causal. The second involves what we *believe* causal relationships to be, i.e. the process of assigning characteristics to relationships that identify them as causal (as opposed to spurious or something else altogether). The two processes are intertwined. However, the first is a matter for psychologists and philosophers and is touched on only briefly in this discussion, in the subsections *Changes in status* and *Time* below. The second lies at the heart of the development of any causal framework and is discussed in more detail in the subsection *Invariance.*

Changes in status

The relationship of cause and effect is perceived as a series of changes in status. For example, consider the statement *My headache is gone because I took an aspirin.* Both *taking the aspirin* and *the headache going away*, are changes in status. The first describes a change brought about by an intentional act, a *decision* or *intervention.* The second is instead a change, perceived as a consequence of the intervention, of a physiological characteristic, headache status.

Now consider the statement *the pavement is wet because it rained.* The "cause" here is the change in status of the weather from not-raining to raining, and the "effect" the change in status of the pavement from dry to wet. In this case, both the cause and the effect are changes in status of natural characteristics of the weather on the one hand and the pavement on the other. Neither is brought about by human intervention.

Two points worth noting, the first is that causes are understood in contrastive terms. We contrast an initial state with a changed state. How I felt before I took the aspirin is compared to how I felt after I took it.

The second point is that as seen in the previous section on the expression of causal concepts, there are two distinct types of changes that can be seen as causes. The first example links cause to human intervention, the second attributes cause to natural events.

Time

The importance of time ordering in our understanding of causal relationships is fundamental. In the methodology proposed in this thesis, time ordering is not mentioned explicitly. However the causal structures used incorporate a natural ordering, and it is assumed that an effect cannot precede its cause in time. This is reflected in the conditional independence constraints and in the graphs that encode them.

Invariance

Although we perceive causal relationships in terms of changes in status, these differ from other types of changes in that we believe them to be invariant or stable (Pearl 2000 and 2001a). The concept of invariance is best explained through examples;

Example 2.1.1

If all the rats in a laboratory that have been administered a certain drug have overcome a disease, whilst all the rats that have been administered a placebo have not, we might say that it is the treatment that *caused* the rats to overcome the disease.

Given this premise, we expect that if we let the rats loose in the wild and somehow they are administered the drug, they will overcome the disease or at least have a higher chance of overcoming the disease than rats that are not administered the drug, if all the important factors have been taken into

consideration. Also we expect the relationship to persist if we change the dosage by small amounts, or administer the drug as an injection or orally. Further, we believe that the relationship will persist if we intervene on it by changing the environment in which it takes place, again given all important factors are taken into consideration.

Thus we believe that the drug will work in the same way in the UK as it does in Mexico.

Example 2.1.2

Another example of invariant relationships are the laws of Physics. For example, consider *Ohm's* law relating voltage, V to current I and resistance R,

$$V = I \times R$$

We believe that it is invariant with respect to interventions on the voltage or resistance of a circuit. So if we changed the battery on a circuit with resistance r from v_1 to v_2, we would expect the current going through the circuit to go from $i_1 = v_1/r$ to $i_2 = v_2/r$, in accordance with the formula above.

As we believe that causal relationships are invariant to interventions, we conversely tend to assume that relationships that are invariant to interventions are causal. In fact in this work we characterise causal relationships in this way.

A relationship we do not believe to be causal, and thus not invariant to interventions is the relationship between the water level in Venice and the price of bread in the UK (Sober 1988) even though both have been rising steadily since records began. The reason is that we believe that if the price of bread in the UK were halved, the water level in Venice would not decrease. However, until we have done this, we will not know for sure.

Example 2.1.1 and 2.1.2 are very different and demonstrate one of the schisms in the field of causal inference. Ohm's law is a law of Physics, and is generally considered to be one of many stable mechanisms that rule the world. The laws of Physics work very well and make predictions that are subject to very little uncertainty. What will happen in the far future can be predicted, and what happened in the remote past can be deduced and explained. These laws seem to be self consistent so far, meaning they do not contradict one another and form a harmonious world view, although there are different fields that deal with the macroscopic and the microscopic. These laws reflect the way the universe works and are generally expressed in terms of deterministic functions.

The subject of example 2.1.1, whether rats are cured by a particular drug, is not so clearly deterministic. In fact, most problems statisticians deal with are of this type, where no precise formula that makes accurate predictions and fits available data well can be easily found. However, most literature in causal inference (with the notable exceptions of Spirtes et al. (2000) who adopt a purely graphical approach) , Pearl (2000), Robins (1986), Heckerman

and Shachter (1995) and Rubin (1974) amongst others, follow the economics and social sciences *Structural Equations Model* (SEM) method, and describe causal relationships in terms of deterministic functions.

Following a SEM approach to analyse example 2.1.1, we say that Y is the response, representing whether the rat has overcome the disease or not, T represents whether the rat has been administered the drug or not and finally some additional covariate information about the rat such as age and sex is given in X. We model the relationship between the response and the explanatory variables in a function as follows:

$$Y = f(T, X) + \epsilon,$$

where ϵ is an error term with an appropriate probability distribution.

Determinism

In my opinion (as well as Dawid (2000),Shafer (1996), Didelez (2003)) it is inappropriate to model every causal relationship as a deterministic function. It is more appropriate to express them as probability distributions, so in the above case I am interested in $p(Y|T, X)$, the probability of a particular response given the treatment and covariate information.

Saying that all causal relationships can be expressed as deterministic functions is an assumption that cannot be verified, and hence I am not willing to make this assumption unless, as in the case of the laws of Physics, the validity of such an assumption has been confirmed beyond reasonable doubt.

Further, I am not sure, and am not willing to assume that the universe is a complicated deterministic machine, where someone with enough information would be able to predict everything.

The concept of invariance to interventions corresponds to how I identify causal relationships. This is as much as I am willing to assume about the nature and existence of causal relationships.

2.1.3 From metaphysics to methodology

There follows a purely metaphysical discussion that relates the aspects of causality described thus far to my choice of methodology. The methodology itself is indifferent to these considerations, and can be used independently of or despite them.

Humans as (potential) agents

When we say *the pavement is wet because it rained*, I believe we mean *the pavement is wet because noone covered it and it happened to rain.* Indeed, if I had parked my car over a patch of pavement, I would have prevented it from being wet even if it had rained.

Contrast this with the statement *John is not pregnant because he is a man.* Do we really consider the fact that *John is a man* a cause of *his not being pregnant*? I think not, in fact, *John is not pregnant because he is a man* is a nonsensical statement. Yet the elements of this statement are the same as those in the statement about the pavement, with the exception that we are able to act in the case of the pavement. It is this that makes the statement

about the pavement make sense.

This leads me to believe that causal relationships can be understood only in terms of the human being as the (potential) *agent.*

Cause as a product of the human mind

The above and the belief that what human beings perceive is not necessarily what happens in the *real world* beyond our senses further leads me to believe that causal relationships are the product of the human mind.

An important consequence of my belief that causal relationships are attributes of our perception of the world, and not necessarily of the real world outside our minds, is that I must be very careful about the assumptions I make relating these two worlds. I must try and make assumptions that relate well and as much as possible to what I can perceive, as this is within my realm of experience, and say as little as possible about what I cannot.

This argument leads me to forgo a deterministic approach to causality, as I am unwilling to make assumptions about how the real world works, and adopt a purely probabilistic approach.

It also leads me to discard the counterfactual approach, as this is based on comparing what happened to what might have happened. A counterfactual is a quantity I will never see, and I am unwilling to make assumptions about *a)* its existence and *b)* its behaviour.

In my opinion, probability itself as a concept is also a product of the human mind. It is our attempt to quantify uncertainty in our perception of

the world around us. As it is my aim to remain as much as possible in the realm of what we can perceive, the use of probabilistic concepts is also very important.

2.2 Framework Development

This section covers the basic building blocks of the decision theoretic causal inference methodology. First I state what requirements a causal inference methodology should fulfil, then go on to explain why statistics is used as the basis of causal inference. Thus, the decision theoretic framework is adopted as a basis for causal inference and augmented by a new semantic suitable for causal inference. This semantic is introduced after the basic notation is laid out. The approach proposed follows Dawid (2000) and (2002).

2.2.1 Requirements

Requirements 2.2.1 A causal inference methodology must:

(i) be meaningful with respect to what we believe cause to be,

(ii) be targeted at how I intend to use the knowledge that I gain from causal relationships and

(iii) make as few untestable assumptions as possible and base inference as much as possible on what can be observed. By untestable I mean assumptions that cannot be verified by observation or experiment.

For a methodology of causality to fulfil these requirements, it must incorporate the concepts of invariance to intervention, it must answer causal queries that are useful and relevant and it must be based on sound assumptions. These three points cannot be treated separately and a compromise must be found that fulfils each as much as possible. Different approaches put emphasis on some points more than others. The methodology proposed in this thesis is driven initially by the second requirement. In my opinion it amply fulfils the other two requirements, however these are the more contested aspects.

2.2.2 Causality and Statistics

Most scientific investigation uses statistical analysis to back up causal hypotheses. It is not rare to read that statistical analysis of the data *proved* that phenomenon A caused phenomenon B. Although such studies cannot be discredited off hand as human intuition usually makes the correct connections and identifies causal relationships, statistical analysis alone is not sufficient to establish causal links.

As a consequence causal inference was largely ignored until the late 70's (Rubin 1974). However, the fact that *correlation does not imply causation* has not stopped statisticians from making indirect references to causality. Many commonly used terms in statistics are in fact causal in nature. Spurious correlation, confounding and even randomisation are concepts that do not make sense without invoking causation. A thorough discussion of this overlap

is given in Pearl (2000) and (2001a).

Although probability theory and statistics alone cannot express causal concepts, they can serve as the basis for a causal inference framework. This is because the information we get from the world is subject to many distortions; measurement error, exceptional events, and when dealing with human beings, their natural unpredictability. Our data is warped and imprecise. The language of probability enables us to quantify this uncertainty, even when we are investigating what we believe to be invariant relationships. The fundamental problem is that in order to make causal inference it is necessary to augment the standard probabilistic vocabulary to include notation for causal concepts. This has been done in the literature, each approach differing to a greater or lesser degree from the standard probabilistic vocabulary.

Exchangeability

An important aspect of statistics that enters in the discussion of causality is that of exchangeability. In order to make causal inference based on statistics, we are faced with the problems that emerge from sampling and data gathering. In particular, we must be willing to assume that the units treated in the past are representative of the units we wish to treat in the future.

Consider a simple example. A pharmaceutical company is interested in the causal effect of taking aspirin on headaches. They run a clinical trial and gather data from which they estimate $p(headache|decision = take\ aspirin)$ and $p(headache|decision = not\ take\ aspirin)$. The question they ask them-

selves is *Given the information we have gathered from the clinical trial, can we say that taking an aspirin will make a potential aspirin buyer's headache go away?* [1]

In order for the data gathered in the trial to be relevant for inference on a potential buyer, the assumption of exchangeability of participants and potential buyers must be made.

Formally the data points $X_1, \ldots, X_n$ are exchangeable in their joint distribution given some parameter θ if $p(X_1, \ldots, X_n|\theta)$ is invariant to permutations of the indices $(1, \ldots, n)$.

As in a clinical trial it is generally assumed that the treatment and control groups are exchangeable, the new unit is generally assumed to be exchangeable with all units in the trial before treatment has been administered.

2.2.3 Decision Theoretic Framework

Given that probability is used as the basis for this causal inference framework, what additional assumptions need to be made in order to make the leap into causality?

Here, the assumption is made that causes are types of interventions for reasons explained in section 1.2. Interventions are types of decisions, so

Assumption 1 *: Causes are decisions.*

It is possible to relax the assumption if the concept of causation is relaxed

[1] We do not consider problems resulting from not blinding , non-compliers or missing data.

and events other than decisions are considered causes. I reiterate that the mathematics is indifferent to how causes are defined.

Is this a reasonable assumption to make? In order to determine this, it is necessary to assess the assumption in the light of the requirements made in list 2.2.1. However, it is difficult to do this without looking in more detail at the basic mathematical framework this assumption leads to. I therefore postpone this discussion to sections 2.2.4 and 2.2.5.

The natural consequence of considering the decision as the basic concept in causality is adopting a decision theoretic approach, such as that developed by Lindley (1985) and Raiffa (1970). Using this set-up, causes can be expressed as decisions, and causal effects as aspects (differences in expectations or utility functions) of probability distributions *conditional on decisions.*

What type of interventions are there? Which do we consider here?

Interventions

In broad terms there are three ways in which a variable or a system of variables can be intervened upon. The first and most simple is the *point intervention.* This intervention consists of forcing a variable to take on a specific value with no uncertainty.

Another type of intervention is a *strategy*, where a point intervention takes place on a variable depending on the value taken by other variables. For example, a doctor may decide that if his patients are over 60 he will always prescribe cod-liver oil, and if they are below 60 he will prescribe cod-liver oil

only if they complain about joint ache. A strategy can be either static or dynamic. In the first case, a predetermined intervention is applied at given points in time, in the second, the intervention depends on the status of the variables at each time point.

We can also consider *randomised interventions*, that is interventions where the value assigned to a variable is not predetermined, but is drawn from an appropriate probability distribution.

Another type of intervention is changing the environment or location of a causal relationship. For example, if we run a clinical trial in the UK and then another in the US, the location will be a factor that is within our control. Although this has not been explicitly covered in this thesis, it is possible to formulate a change in location as a type of intervention variable.

Finally, there are *black box interventions* where we *kick* a closed system. A typical example of such an intervention is the introduction of a policy, such as the congestion charge in London, where cars are allowed into the city centre only if they pay a 5 pound congestion charge. The traffic system changes due to this policy intervention and settles into a new equilibrium. However, black box interventions are very complex and I suspect that extra structure would have to be added in order to tackle the expression and estimation of causal effects.

This thesis will principally consider point interventions, although random interventions are developed in Chapter 5. The theory could easily be extended to include strategies, but these will not be discussed. See Dawid (2002) for

some extensions.

Treatment definition

Treatments must be well defined. Even in the case where the treatments are not physically possible (Heckerman and Shachter's framework described in section 3.4 does allow causes to be variables that cannot physically be intervened upon, such as sex) it is important to define these clearly and the specific context in which they are being considered.

Also, when considering an experimental situation, a change in design might have an effect on the quantities we are trying to estimate. An extreme example is using observational studies to make causal inference. The fact that in an experiment a variable is a decision, and in an observational study it is not, changes the nature of the estimates. See Chapter 4 for an in-depth discussion of this issue.

Causal Variables

In this section I define causal variables in the decision theoretic conception. Note that causes are defined contrastively, that is, whenever we say a decision causes something, it is relative to another decision.

Definition 2.2.1 (Causal Variable) *Let $I_D = d$ for variable D be shorthand notation for intervening on D such that $D = d$. The variable D is a causal variable with respect to X (or just a cause of X) if the probability distribution of X conditional on $I_D = d$ is not the same as the distribution of*

X conditional on $I_D = d'$ for some $d, d' \in \mathcal{D}$ the domain of D, where $d \neq d'$. That is

$$p(X|I_D = d) \neq p(X|I_D = d').$$

I_D can be any type of intervention or manipulation of D. In particular, this means that a variable that *cannot* be intervened upon, such as the sex of a person or their age is not a cause in this conception.

Causal Effect

To estimate the effect of an intervention on a specific individual we adopt a Bayesian approach and use the predictive probability of response given the intervention. Consider again the example where a pharmaceutical company has run a clinical trial to determine the causal effect of aspirin on headache status. The patients were assigned at random to take aspirin or not when they had a headache. The aim now is to estimate the causal effect of an aspirin will have on a new individual, given the exchangeability assumption. To do this we use the data gathered in the clinical trial and assign appropriate prior distributions and calculate the predictive probabilities in the standard Bayesian way. From the predictive distributions we calculate the difference of the expectations or other functions of interest.

In this thesis we will be using the expectations recovered directly from the data, and not full predictive distributions for the sake of simplicity. The difference in expectations of the response given two different treatments, is

the *average causal effect* (ACE) and is generally the causal quantity we seek to estimate.

For example, we might be interested in the effect of treatment $I_D = 1$ on response X relative to the effect of treatment $I_D = 0$ on response; the causal quantity would then be the average causal effect of $I_D = 1$ with respect to $I_D = 0$ given by

$$E(X|I_D = 1) - E(X|I_D = 0). \tag{2.1}$$

This is not to be confused with $E(X|D = 1) - E(X|D = 0)$ which is the difference of expectations given that the values of D arise naturally, i.e that we happen to observe X given both $D = 1$ and $D = 0$.

We saw in the last section that it is important to make the assumption of exchangeability in order to make causal inference. In terms of the decision theoretic framework this means:

Assumption 2 : *The new unit for whom we want to infer the causal effect of one decision with respect to another is exchangeable with the units used to estimate the probability of response given the two decisions* ***before*** *any decision has been made. In particular, the new unit must be exchangeable with the units who were administered the same treatment as it will be administered.*

2.2.4 Asking a sensible question

Of the three requirements in list 2.2.1, it is item *(ii)*, that drives the adoption of the decision theoretic framework as the basis for causal inference. The idea behind this requirement is that we do not ask causal questions that are not useful. A useful question is a question whose answer can provide us with relevant information for future decisions, whilst making as few untestable assumptions as possible. It is my opinion amongst others (see Dawid (2000),Shafer (1996), Didelez (2003)) that causal questions that lead to counterfactual formulations in particular are not useful questions, in fact they can be positively misleading.

Questions that are not useful are of the type *would my headache have gone away if I had taken an aspirin?* or *would he have a higher paid job if he had gone to university?*. In order to answer the former, it would be necessary to know what would have happened to me had I taken the aspirin, a state of the world that did not come to be, and which I have no empirical information about. Thus these questions are not useful because they cannot be answered without making strong assumptions about events that did not happen, and further, provide us with no more information (even if they could be simply answered) than the decision theoretic formulation: *given clinical trial data on the effect of aspirin on others like me will my headache go away if I take an aspirin?* which can be answered relatively simply given appropriate data.

Further, the aim of the inference in both cases is to inform future decision making processes: in the first case, whether I should take aspirin the next

time I have a headache, and in the second whether a young person should be encouraged to go to university with the prospect of attaining a better paid job. Both can be reformulated in terms that makes them easy to deal with in terms of the decision theoretic framework. The first question becomes: *given clinical trial data on the effect of aspirin on others like me will my headache go away if I take an aspirin?* and the second becomes, *given data on the general population will he get a higher paid job if he goes to university?* In terms of the decision theoretic formulation, the answers to these questions depend only on the marginal distributions of responses conditional on decisions, both of which can in principle be estimated from data. No comparisons with events that did not happen need to be made.

In fact, the only questions that are useful are those that are of the form *given clinical trial data on the effect of aspirin on others like me will my headache go away if I take an aspirin?.*

This statement is bold. In particular if we consider expression *4)* in list 2.1.1, we see that the question being addressed here *is* useful, and apparently has to be formulated in a counterfactual way.

In a court of Law, if the claimant says he was discriminated on the grounds of his race, the question of interest is whether the defendant *would have acted differently had* ***this particular claimant*** *not been black.* However, it would suffice to determine whether the hirer is racist to win the case for the claimant. Hence formulating the question as *given past hiring practice of this and other companies, will this company hire someone if they are black?* would be just as

good, given the relevant covariate information about whether they are indeed suitable for the job.

This is a slightly different formulation and begs the question *why should this particular claimant be compensated?*. The answer is that this claimant brought up the problem and if the company discriminates, has probably suffered from it and should thus be compensated.

A further consequence of this type of trial is the introduction of anti-discrimination laws in the future, thus ultimately, the trial aids future decision making processes.

Retrospective questions

Shafer, in his comment to Dawid (2000) claims that retrospective questions, like counterfactual questions, are "silly". For example, say I know for certain that if I take a glass of water and two aspirins the headache goes away, and if I take less water or less aspirin the headache stays. Then if I take the glass of water and the two aspirins and ask question *My headache is gone, is it because she took the aspirin?* I cannot answered this with a simple yes or no although the question itself is perfectly understood. This is because the aspirin is just one of many possible causes for my headache having gone.

Retrospective attribution of causality cannot be dealt with in the decision theoretic framework as it stands. No attempt is made at tackling this problem in this thesis, and all inference is limited to prospective causality. This is in accordance to the idea that we use causal inference to inform future decision

making processes.

2.2.5 Requirements revisited

Given that the choice of methodology is driven by requirement *(ii)* in list 2.2.1, does the decision theoretic approach fulfill the other two requirements?

The first requirement states that the framework must reflect what we believe cause to be. As I believe that cause is determined by human actions, I have introduced this by assuming causes are decisions.

The idea that causal relationships are invariant to interventions is reflected in the methodology as follows. Consider the ACE (2.1). As it is a causal quantity, and thus invariant to interventions, we assume that it can be used to make inference about other units in the future given these are similar enough even if the intervention is slightly different.

The third requirement is fulfilled in my opinion by only making assumptions 1 (page 34) and 2 (page 39). The first assumption is a very strong one. However, it does reflect what we believe causes to be, and there is a general consensus in the literature that causes should be seen in terms of interventions, although this is not always made explicit. Also, no additional untestable assumptions about the existence of counterfactual variables or about causal relationships being deterministic functions are made.

The assumption of exchangeability is usually made in statistical analyses and is thus made in the causal inference literature. It is worth noting however, that whether a set of units is considered exchangeable with another is a

somewhat subjective judgement. It will depend on the information available and on the particular question at hand. It is up to the analyst to determine whether the assumption is plausible. It might also be possible in specific circumstances to relax the exchangeability assumption to one of partial exchangeability. However this is not discussed in this thesis.

Also, the probabilities at the basis of this framework are estimated from available data when possible. In fact I try to adhere to de Finetti's observability criterion which states that *it is legitimate to assess a probability distribution for a quantity Y only if Y is observable at least in principle* Dawid (2000) G.Shafer comment.

Finding causal relationships

Although in Chapter 4, we discuss the assumptions underlying a process that purports to find causal relationships from observational data, it is not the aim of this thesis, nor the purpose of developing this methodology, to *find* causal relationships. The methodology is to be applied when a particular causal relationship is assumed or known to exist between variables.

2.3 Causal Semantics

In the second part of this chapter, the causal semantics that form the basis of the proposed causal inference framework are developed. The graphical models associated to them will be developed in section 2.4.

We start by translating the metaphysical and heuristic arguments dis-

cussed in the first part into formal mathematics.

Before developing the causal semantics formally, we define the concept of conditional independence in section 2.3.1, as it is a key concept in this thesis. Then *intervention variables* are defined with the help of two simple examples. These will also aid in the understanding of causal notation introduced in this section. Interventions are the key to transforming the purely probabilistic into the causal. The algebra is then developed and some important issues, conditioning on interventions or observation and the idea of regimes are discussed. Finally causal effects are defined.

Following Assumption 1, which states that causes are interventions, we make the jump from the purely probabilistic to the causal by introducing a variable that indicates whether we are interested in a probability distribution subject to *natural conditioning* or a probability distribution *conditioning on an intervention.* The approach taken in this thesis follows Dawid (2002).

Some Basic Notation

We denote by upper case letters X, Y A, etc. random variables and/or nodes in a DAG. Sets or collections of random variables are denoted by bold face upper case letters. Thus $\mathbf{V}$ denotes a set of random variables. Realisations or instances of a random variables are denoted by lower case letters, thus if X is a random variable, then x is a possible realisation of X. Further the domain of X is given by the upper case cursive letter $\mathcal{X}$, so that $x \in \mathcal{X}$. Finally, the realisation of a set or collection of random variables is denoted by lower case

bold face letters. So if $\mathbf{U} \subseteq \mathbf{V}$ is a subset of a set of random variables $\mathbf{V}$, then a realisation of every member of $\mathbf{U}$ is denoted by $\mathbf{u}$.

2.3.1 Conditional Independence

Let A and B be two random variables. Denote by $p(A, B)$ the joint probability distribution of A and B, $p(A)$ the marginal distribution of A and, $p(B|A)$, the conditional probability distribution of B given A.

Consider a set of random variables $\mathbf{V}$. Let A, B and C be variables in $\mathbf{V}$.

Definition 2.3.1 *We say that A and B are **marginally independent** if $p(A, B) = p(A)p(B)$ and write $A \perp\!\!\!\perp B$ (Dawid 1979). We say that A is **conditionally independent of B given C** if $p(A, B|C) = p(A|C)p(B|C)$ or equivalently, $p(A|B, C) = p(A|C)$. We write $A \perp\!\!\!\perp B|C$.*

The conditional independence statement $A \perp\!\!\!\perp B|C$ and its properties are explored in detail in Dawid (1979).

2.3.2 Intervention variables

To facilitate the understanding of intervention variables, there follow two examples.

Example 2.3.1 *Aspirin clinical trial*

An pharmaceutical company is running a clinical trial on the effectiveness of aspirin. The participants are believed to be representative of the general population.

When a participant suffers from headache, he or she is randomly administered an aspirin or not. [2]

Let the variable that indicates whether treatment was received be T. T is 1 if an aspirin is received and 0 otherwise.

The response to treatment is denoted by Y. The response is also binary, taking on value 0 if the headache does not go away within two hours of receiving treatment and 1 if it does.

Further, let X be a known covariate. These are generally patient characteristics. Initially let $X = (S, A)$ be the vector containing the sex S and age A.

We assume that each unit receives only one treatment during a given study unless otherwise specified.

Example 2.3.2 *Aspirin survey*

Another pharmaceutical company is interested in what the *natural* intake of aspirin is in the general population. They commission a survey which asks each respondent whether they take aspirin T, whether their headaches go away after 2 hours Y when they do and don't and finally, what their age A and sex S are. The domains of the variables are the same as in the previous example. The respondents are also assumed to be exchangeable with the same general population.

[2]For the sake of simplicity, we do not take into account any ethical concerns. Or the problem of the participant knowing whether he/she is receiving treatment.

As discussed in section 2.2.3, interventions can take on many forms. This thesis mainly uses point interventions, and it is by defining these that we are able to easily express and distinguish between the concepts of conditioning by intervention and natural intervention.

Definition 2.3.2 *For a variable T define the* ***intervention variable*** *on T, denoted by F_T as follows; $p(T = t|F_T = t) = 1$, where $t \in \mathcal{T}$, the domain of T. That is T is forced to take on the value t if $F_T = x$. If $F_T = \emptyset$ then T arises naturally, and we say that F_T is idle.*

The intervention variable is not a random variable. It is a type of decision variable (Dawid 2002) and there is no marginal distribution associated to it.

Definition 2.3.3 *Define* ***natural conditioning*** *as conditioning on $(T = t, F_T = \emptyset)$, and* ***conditioning on intervention*** *as conditioning on $(T = t, F_T = t)$.*

The two types of conditioning lead to different probability distributions generally. To see this, consider examples 2.3.2 and 2.3.1. The joint distribution of the variables in the problem is $p(Y, T, X|F_T)$, where F_T is the intervention variable on T. In example 2.3.1, this joint distribution is given by $p(Y, X, T = t|F_T = t)$, as we intervene on the system that relates aspirin intake and headaches by forcing participants to take or not aspirin. Hence $F_T \neq \emptyset$. In example 2.3.2, there is no external intervention on the system, and hence $F_T = \emptyset$, and the joint probability distribution is given by $p(Y, X, T|F_T = \emptyset)$.

When we consider the problem we see that in the survey, we cannot exclude the possibility of T, aspirin intake, depending on X, the respondent's age and sex. In the clinical trial however, as the treatment is randomised, there is no dependence between X and the treatment T, thus $X \perp\!\!\!\perp T | F_T \neq \emptyset$. Further, we assume that X and F_T do not depend on each other marginally. This generally makes sense in randomised trial when $F_T \neq \emptyset$. Given this consideration, we look in more detail at the joint probability distributions;

Example 2.3.2 (Survey):

$$\begin{aligned} p(Y, X, T = t | F_T = \emptyset) & \\ &= p(Y|X, T = t, F_T = \emptyset) p(X|T = t, F_T = \emptyset) \\ &\times \; p(T = t | F_T = \emptyset) \end{aligned} \tag{2.2}$$

Example 2.3.1 (Clinical Trial):

$$\begin{aligned} p(Y, X, T = t | F_T = t) &= p(Y|X, T = t, F_T = t) p(X|T = t, F_T = t) \\ &\times \; p(T = t | F_T = t) \\ &= p(Y|X, T = t, F_T = t) p(X|F_T = t) \qquad (2.3) \\ &= p(Y|X, T = t, F_T = t) p(X) \qquad (2.4) \end{aligned}$$

The case for example 2.3.2 is straightforward, we have chosen this factorisation of the joint probability distribution for the sake of comparison with example 2.3.1. In the case of example 2.3.1, the same factorisation has a different expression. First $p(X|F_T, T)$ is equal to $p(X)$ by the conditional independence argument given above, and second, as there is no uncertainty

about the value of T when it is set $p(T = t|F_T = t)$ is 1, and disappears from the formula.

Consider $p(Y|X, T = 0, F_T = \emptyset)$, the distribution of the response given that no aspirin was taken ($T = 0$) by a respondent in the survey ($F_T = \emptyset$). This is *not* the same as $p(Y|X, T = 0, F_T = 0)$, the probability of the response given that no aspirin was administered ($T = 0$) to the participant in the clinical trial ($F_T \neq \emptyset$). This is because an individual who is participating in a clinical trial is not guaranteed to respond to not receiving any treatment in the same way as a survey respondent who chooses not to take aspirin if he or she has a headache.

We use the expression intervention variable and point intervention to mean the same thing for the remainder of this thesis, as we deal mainly in point interventions.

Observation vs Intervention

Two themes that run through this thesis are the relationship between observational and experimental data and the problem of identifying causal quantities, that is functions of probabilities conditional on interventions from observational data.

By *observational* data we mean data that has been gathered passively from studies where no interventions have taken place, such as for example surveys. We say that such data is subject to natural conditioning. By experimental data, we mean data gathered in an experimental setting where controlled

interventions or manipulations have taken place. In very simple terms, the problem of identification of causal effects from observational data is that we wish to make inference about $p(Y|T, F_T = t)$ from $p(Y|T = t, F_T = \emptyset)$.

In terms of the above example this is analogous to saying that we wish to make inference about the response to aspirin in the clinical trial from the survey data. If certain conditional independence relationships hold, or equivalently, we are willing to makes strong assumptions about how experimental and natural conditions are related to one another, then it is possible to make inference about the trial using the survey data.

The problem of identifiability is covered in more detail in subsection *Identifiability* in section 2.3.3 , and aspects of it are covered in Chapters 5 and 6.

Regimes

We have seen how the intervention variable is defined. We also assume that when we seek to determine cause, we are interested in comparing the probability distribution of a response variable conditional on one intervention, i.e. $p(Y|T = t, F_T = t)$ with the probability distribution of a response variable conditional on another intervention $p(Y|T = t', F_T = t')$ say. Further, as has been touched on briefly in section 2.3.2 *Observation vs intervention*, we are sometimes interested in evaluating causal effects i.e, when $F_T \neq \emptyset$, using data recorded under natural conditioning i.e, when $F_T = \emptyset$. These settings of F_T are examples of *regimes* of T. The regimes of T are thus all possible types of

interventions on T, and the idle setting, denoted by the idle regime.

2.3.3 Causal Effects

We will consider various types of causal effects in this thesis. The simplest of these, and the basis for the others, is the average causal effect (ACE) already mentioned in the initial part of this chapter in (2.1).

As discussed in section 2.2.3, causal effects are always measured comparatively. That is by comparing the response of one intervention to that of another. Hence we formally define a causal effect of one intervention relative to another as follows:

Definition 2.3.4 *Let X and Y be two variables, such that X can be intervened upon with intervention variable F_X. Let Y be a characteristic of interest of a set of units $u \in U$. The* ***average causal effect*** *(ACE) of $F_X = x$ relative to $F_X = x'$ on Y is given by*

$$E(Y|F_X = x) - E(Y|F_X = x'). \tag{2.5}$$

If we want to use the ACE for inference on a unit u_{new}, which is usually the case, we must assume that u_{new} is exchangeable with the units in U. Therefore inference made on them is valid for u_{new}.

Identifiability

A typical problem in causal inference is to determine when it is possible to evaluate causal quantities, that is functions of probabilities conditional on

interventions, from data that are observational and hence involve no interventions. In order to be able to identify causal quantities from observational data, it is often necessary to make additional assumptions. Some of those adopted in counterfactual frameworks are discussed in section 6.6. In this section we look at identifiability without making additional parametric or modelling assumptions.

It is worth noting that the concept of exchangeability is very important to determine whether a quantity is identifiable. This is particularly clear when making causal inference from observational data. Are the subjects in the observational study exchangeable with the new units of inference who will be intervened upon? This assumption is generally made, whether it is appropriate tends to be context specific.

We define identifiability of causal quantities *in principle* and *in practice.* A causal quantity is *identifiable in principle* if it is *physically* possible to perform an experiment that will allow us to compute it.

For example, the effect of aspirin on headache is identifiable in principle as it is physically possible to perform a clinical trial that assigns aspirin to one group and a placebo to another, and records the persistence of headache in the two groups. Also, the effect of smoking on lung cancer is identifiable in principle as it is physically possible to perform a trial where half the participants were forced to smoke and the other half were forced not to, and the incidence of lung cancer recorded in both groups. Clearly, such an experiment is not performed on humans (although animals have been experimented on)

because it is un ethical, however, it is possible. The individual causal effect in counterfactual terms, that is the difference in response to treatment and placebo for a specific unit at a specific time, is not identifiable in principle, as we cannot imagine an experiment that involves a patient receiving both treatments and responding to them simultaneously as though each were the only treatment received.

A causal quantity is *identifiable in practice* in two distinct situations. The first is if the causal quantity is identifiable in principle and the trial has been performed. The second situation in which a causal quantity is identifiable in practice is when it is possible to manipulate the assumed set of conditional independences to reformulate an unobserved causal quantity (generally a function probabilities conditional on interventions) in terms of observed quantities. These can be probabilities estimated from observational data, probabilities estimated from experiment or even a combination of the two. It is interesting that in the latter type of identifiability in practice, it is not necessary for the quantity to be identifiable in principle. An example of this is a component of the effect of the treatment on the treated discussed in Chapter 6. No experiment can be imagined that will make it possible to estimate this quantity, however, it can be derived from formulae in which all other parts are identifiable in principle.

2.4 Graphical Concepts and Notation

In this section the graphical model notation that will be used for the remainder of this thesis is defined. We start by defining graphs, and then go on to define *directed acyclic graphs* (DAGs). We describe how these can be used to model probabilistic structure. Then we see that DAGs lend themselves to causal interpretations, but that additional structure must be introduced in order to do this unambiguously. Thus we introduce *augmented DAGs.*

2.4.1 Conditional Independence Graphs

The graphical notation is based on Lauritzen (2001).

Definition 2.4.1 (Graph) *A* ***Graph*** *is a pair* $\mathcal{G} = (N, E)$*, where* N *is a set of nodes and* E *is a subset of* $N \times N$ *of ordered pairs of nodes called edges. We require that there are no multiple edges, which is fulfilled as* E *is a set, and further, that the elements of* E *consist of distinct nodes, so that there are no edges going from a node to itself in a loop.*

If A and B are two nodes in N we say that there is a *directed edge* from A to B if $(A, B) \in E$ $A \rightarrow B$. When all the edges in a graph are directed, the graph is called a *directed graph.* The ordering of the nodes indicates the direction of the arrow, thus $(A, B) \equiv A \rightarrow B$, however, $(B, A) \equiv B \rightarrow A \neq (A, B)$.

A *path* of length n (where $n > 0$), from A to B in the directed graph $\mathcal{G}$, is a sequence of distinct nodes $A_0, A_1, \ldots, A_{n-1}, A_n$ such that $(A_{i-1}, A_i) \in E$ for all $i = 0, \ldots, n$, where $A = A_0$ and $A_n = B$. As the nodes are distinct, the

path never crosses itself and as the graph is directed, the path always follows the direction of the arrows.

A *cycle* in a directed graph is a path with the difference that the first and the last nodes are the same.

Definition 2.4.2 (Directed Acyclic Graph) *A DAG is a graph* $\mathcal{D} = (N, E)$ *such that*

1. $\mathcal{D}$ *is a directed graph and*

2. $\mathcal{D}$ *contains no cycles.*

This thesis will only use DAGs.

A node $Pa(A) \in \mathcal{D}$ is a *parent* of a node A if $(Pa(A), A) \in E$. A node $Ch(A)$ is a *child* of node A if $(A, Ch(A)) \in E$. In a DAG, a variable B is said to be an *ancestor* of A if there is a path from B to A. The set of ancestors of A is denoted by $an(A)$. Similarly, A is a *descendant* of B, and the set of descendants of B is denoted by $de(B)$.

Definition 2.4.3 (Moral graph) *The* ***moral*** *graph* G^m *for the graph* G *is the undirected graph made from* G *by first joining with an undirected edge, all parents of a common child that are not already joined, and then creating the undirected version of this graph. Transforming a graph into its moral version is termed* ***moralisation***

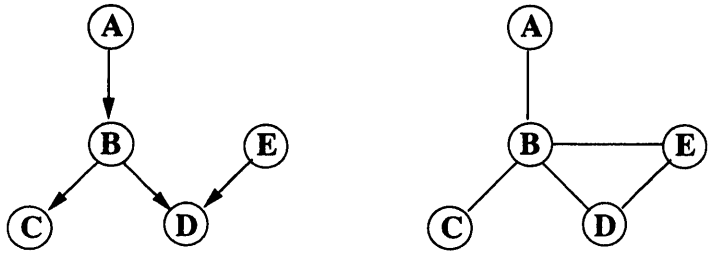

Figure 2.1: The undirected graph on the right is the moralised version of the DAG.

DAGs and Conditional Independence

Consider the DAG in figure 2.1. It encodes the following set of conditional independences:

$$(A, C) \perp\!\!\!\perp (D, E)|B$$
$$A \perp\!\!\!\perp C|B$$
$$E \perp\!\!\!\perp (A, B).$$

The conditional independences can be *read off* the graph using the *moralisation* criterion (Lauritzen 1996). The moralisation is described in definition 2.4.3 above. The procedure to read conditional independences from graphs will be used often throughout the thesis, thus it will be described in detail for DAG 2.1.

We want to know what conditional independences, if any, exist between A, B and C as well as those between B and E, and B,C and D by reading them off the graph. The method is as follows;

1. First build the *ancestral graph* by excluding all variables that are not

ancestors of the variables of interest or the variables themselves.

ABC Ancestral graph of $A \cup B \cup C$ is $A \rightarrow B \rightarrow C$.

BE Ancestral graph of $B \cup E$ are the disjoint $A \rightarrow B$ and E.

BCD Ancestral graph of $B \cup C \cup D$ is the whole DAG on the left of figure 2.1.

2. Then *moralise* as described in definition 2.4.3.

ABC Moralised and undirected ancestral graph is $A - B - C$.

BE Moralised and undirected ancestral graph is $A - B$ and E

BCD Moralised and undirected ancestral graph is the undirected graph on the right of figure 2.1. B and E are joined as they are both parents of D.

3. Now we investigate the paths between the variables in the moralised and undirected ancestral graph.

ABC All paths from A to C go through B, so we can say that $A \perp\!\!\!\perp C|B$.

BE There are no paths from B to E, so we can say that $B \perp\!\!\!\perp E$, that is B is marginally independent of E.

BCD All paths from C to D go through B, so $C \perp\!\!\!\perp D|B$.

Definition 2.4.4 (Markov Equivalence) *When two or more DAGs encode the same set of conditional independence relationships, we say that they are* ***Markov Equivalent.***

Markov equivalent DAGs are characterised as having the same *skeleton*, and the same *v-structures* (see A). The skeleton is the undirected version of the DAG, the v-structures are sets of three nodes, say A, B and C such that $A \rightarrow B \leftarrow C$ and further, there is no edge between A and C.

The DAG on the left in figure 2.1 encodes the same conditional independence relationships as the DAGs in figure 2.2. It is easy to see that they have the same moralised graph as they have the same skeleton and the same v-structures, in fact there is only one v-structure, $B \rightarrow D \leftarrow E$.

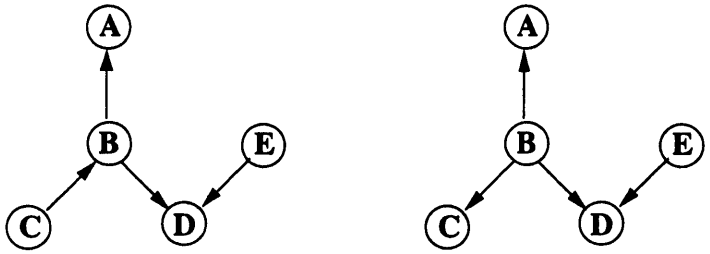

Figure 2.2: These DAGs are Markov equivalent to DAG in figure 2.1.

We have seen how DAGs can embody conditional independence structures. They can also induce factorisations of the joint probability distributions of the variables they represent. For example, the DAG on the left hand side of figure 2.1 corresponds to the following factorisation of the variables A, B, C, D and E.

$$p(A, B, C, D, E) = p(A)p(B|A)p(C|B)p(D|B, E)p(E).$$

The DAGs in figure 2.2, correspond to a different but equivalent factorisation of the joint probability distribution of the variables. For example, the

DAG on the left hand side of figure 2.2 represents the following factorisation:

$$p(A, B, C, D, E) = p(A|B)p(B|C)p(C)p(D|B, E)p(E),$$

while the DAG on the right hand side of figure 2.2 corresponds to

$$p(A, B, C, D, E) = p(B)p(A|B)p(C|B)p(D|B, E)p(E).$$

Note that the part of the factorisation that corresponds to the relationship between D,E and B remains unaltered, as it does in the DAGs, this is because of the v-structure between the three variables.

2.4.2 Graphs to represent causal relationships

DAGs lend themselves to being interpreted causally as they permit no cycles or loops, which corresponds to the idea that cause flows in one direction from causes to effects.

Two variables A and B joined by a directed edge going from A to B can intuitively be interpreted as A *is a cause of* B. However, using DAGs to formally represent causal relationships is not as simple as it may appear.

Initially, we look at the problem without formality in the following example.

Example 2.4.1 *Fake Tan*

Consider the two DAGs in figure 2.3. L is the variable **lamp**, taking on value 1 if an individual went to the tanning salon and tanned under the UV lamp and 0 otherwise. T is the variable **tan**, taking on value 1 if the individual is tanned and 0 otherwise.

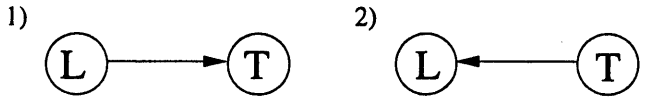

Figure 2.3: DAG 1) makes sense when interpreted causally, whereas DAG 2) does not.

If we interpret DAG 1) causally, it says *lamp (partly) causes tan*, which is in fact the case. DAG 2) however makes no sense when interpreted causally, as it says *tan (partly) causes lamp.*

However, the two DAGs are Markov equivalent, that is, they embody the same set of conditional independences (in this case none as $T \not\perp\!\!\!\perp L$)

This is because L is a decision variable (whether to go to the tanning salon or not), and this is not made explicit in the graph. The solution to this problem is the introduction of the intervention node as described in the next section.

2.4.3 Augmented DAGs

We have seen how to express cause by augmenting probabilities by introducing the intervention variable in section 2.3. We have also seen how DAGs can be used to express probabilistic structure. As the intervention variable can be treated like a variable in the algebraic part of the framework, it follows that it can be expressed as a node in the graphical representation. It is a *decision* node, that is a non-random node and as such is in a box, and not a circle like the random variables in the problem.

Consider the fake tan example 2.4.1. We change the notation to reflect the

introduction of an intervention and hence a causal element to the problem. We replace variable L with variables UV and F_{UV} as follows. We denote by UV the random variable representing exposure to more than a critical amount of ultra-violet radiation. Exposure is given $UV = 1$ and lack of exposure is given by $UV = 0$. Denote by F_{UV} the intervention variable on UV. When $F_{UV} = 1$, the exposure is forced by going under the lamp in the tanning salon. When $F_{UV} = 0$ the lack of exposure is forced, say by choosing not to go under the lamp or staying indoors all year round. When F_{UV} is idle, the exposure is not forced, and UV arises naturally.

We can now reconsider the DAGs in figure 2.3, as augmented DAGs in figure 2.4. DAG 1) now makes causal sense. We can interpret it as meaning

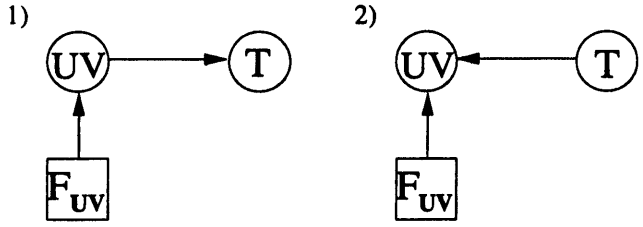

Figure 2.4: DAG 1) represents the causal relationship between exposure to UV radiation and tanning. DAG 2) Still makes no causal sense, however it is no longer Markov equivalent to DAG 1).

that an intervention at UV, say going to the tanning salon and lying under a lamp causes tanning. It is also no longer Markov equivalent to DAG 2), as DAG 1) embodies the conditional independence $T \perp\!\!\!\perp F_{UV} | UV$ where DAG 2) embodies $T \perp\!\!\!\perp F_{UV}$.

Thus we say that an **augmented DAG** is a DAG that contains interven-

tion nodes.

Given an augmented DAG G, we say that the **core** DAG of G to be the DAG left over when all intervention nodes are removed.

2.4.4 Algebra and DAGs

It is clear from section 2.3 that the algebraic part of the methodology does not depend on the graphical part described here, although the opposite is not true. However, the graphical component complements and enhances the methodology, as it makes both visualising the problems and manipulating the conditional independences very easy as we shall see later on in the thesis.

2.5 Conclusion

The setup described thus far is flexible and simple to understand. Any problem of predictive causal inference can be formulated in this model and it can thus be used to make inference for different types of interventions. In particular, the introduction of the intervention node opens the door to nodes which encode more complex manipulations, such as the randomised manipulation node defined in Chapter 5.

As mentioned in the introduction, there are four recurring themes in this thesis. We have seen the assumptions underlying the decision theoretic framework and in the next chapter we will see the assumptions underlying competing frameworks. Also, we have seen that probabilities estimated from data gathered from passive observation or experiment are expressed differently

in terms of the intervention variable. This difference is further discussed in Chapter 4. Chapters 5 and 6 are examples of converting counterfactual arguments into decision theoretic ones, and focus heavily on problems of identification.

Chapter 3

Competing Causal Frameworks

3.1 Introduction

This chapter looks briefly at the competing causal inference frameworks in the literature. In particular, it focuses on the Rubin causal model in section 3.2, Pearl's functional/graphical model in section 3.3 as well as a brief look at Robins' approach to time varying treatments in section 3.3.6. Heckerman and Shachter's model, which extends Pearl's model, is also covered. Finally some comments on the assumptions underlying the models is made and contrasted to the decision-theoretic model proposed in chapter 2.

If the reader is familiar with the causal models in the literature, then he or she can skip the first 3 sections in this chapter and just read section 3.5 where underlying assumptions in the different frameworks are discussed

3.2 Rubin potential outcome framework

This section describes the Rubin Causal Model. Although it was first proposed by Neyman (1923), it was fully developed by Rubin starting in Rubin (1974). The basis for his model is the idea that *..the causal effect of one treatment relative to another for a particular unit is the difference between the result if the unit had been exposed to the first treatment and the result if, instead, the unit had been exposed to the second treatment* (Rubin 1974).

Rubin was the first statistician to explicitly look at the problem of causal inference since Neyman. His approach, based on his field of expertise, Bayesian analysis in the presence of missing data, sparked interest in causal inference which has increased since his 1974 paper.

In the following review of his work, based on Rubin (1978), focus is on how *potential responses* are defined and used as the basis for estimating causal effects where the unrealised response is regarded as missing data. Further, the concept of *ignorability*, fundamental for the identification of causal effects from both experimental and observational data in this framework is considered in some detail.

Section 3.2.1 covers the potential outcomes notation with the aid of an example, section 3.2.2 looks at causal effects in Rubin's model. Section 3.2.3 tackles some of the most important assumptions underlying the model and finally, sections 3.2.4 and 3.2.5 review Rubin's missing data methods applied to causal inference.

3.2.1 Potential outcome notation

In order to explain this framework it is necessary to introduce some basic notation. Following on from the aspirin example 2.3.1, we define *treatment assignment*, the *potential outcome response variables* and the *covariate variable.*

Treatment, Responses and Covariates

First, we need to extend example 2.3.1 as follows: The covariate $X = (S, A)$ is such that $S = 0$ if the unit is a male and 1 otherwise, and A an ordinal variable with three settings indicating age ranges, 0 is < 20, 1 is 20 to 40 and 2 is over 40.

Say there are n participants u_i, $i = 1, \ldots, n$, drawn from a homogeneous population. The treatment received by patient u_i for $i = 1, \ldots, n$ is denoted by T_i. Similarly, the response and the covariate are denoted by Y_i and X_i.

Consider the case of a generic unit. T is the treatment assignment variable. It takes on values $t = 0, 1$. In the example, this corresponds to being assigned a placebo or an aspirin respectively.

In this framework, each unit has two potential responses, one corresponding to each treatment. These are given by Y^0 and Y^1 for treatments 0 and 1 respectively.

The covariates in the example are given by S and A. For the sake of simplicity, say that if one the covariates is observed, then both are. Let X denote the pair of covariate random variables (S, A). For the i^{th} unit in a

sample of n, add the subscript i to the unit so that u_i for $i = 1, \ldots, n$.

Example 3.2.1 *There are 5 patients with headaches in a clinical trial trying to determine the effectiveness of aspirin. Their age and sex are recorded, if possible, (as in example 2.3.1) before they are administered a treatment. Below is a table showing the covariate data, the treatment assignments and the observed responses. Thus patient 5, u_5 is a woman over 40 years of age.*

	Covariate		Treatment	Response	
	A	S	T	Y^1	Y^0
u_1	0	0	0		1
u_2	1	0	1	0	
u_3			0		0
u_4	2	1	1	1	
u_5	1	1	1	1	

Table 3.1: Example: 5 patients in a trial to determine the effectiveness of aspirin.

She was administered an aspirin and after one hour, her headache was gone. Patient 3, u_3 did not have its covariate values recorded.

3.2.2 Causal Effects

Given the above setup, the *individual causal effect* of taking aspirin relative to not taking aspirin for unit i is given by

$$Y_i^1 - Y_i^0. \tag{3.1}$$

In terms of the running example, this means that the individual causal effect for patient 3 is $Y^1 - 0$.

The *average causal effect* is given by

$$E(Y_i^1 - Y_i^0) \equiv E(Y_i^1) - E(Y_i^0). \tag{3.2}$$

This quantity *can* be estimated from the experimental data like the one in the example, without recurring to untestable assumptions such as assumption 3 below.

3.2.3 Assumptions needed to estimate causal effects

In order to estimate the individual causal effect given by (3.1) it is necessary to be able to estimate the values of Y^1 and Y^0. This gives rise to two problems.

Existence of potential responses

The first problem is whether it makes sense to consider the joint distribution or a function of two variables that cannot physically exist simultaneously such as Y^1 and Y^0. Rubin assumes that this is indeed plausible.

Assumption 3 Simultaneous existence of potential responses

The values of the potential responses to all the possible treatments exist for every unit, irrespective of which treatment is administered. The treatment assignment only determines which of the potential responses is observed.

In terms of example 2.3.1, this assumption means that the value of the outcome to both receiving aspirin and receiving a placebo exists for each unit regardless of which treatment the unit actually receives. That is, the values that are not filled in in table 3.1 are assumed to exist even though they have not been observed.

As a consequence of making assumption 3, it is necessary to make an additional assumption, namely, that the value of the potential response given treatment $T = t$, $Y^t = y$ is in fact the value of the response given $T = t$ is the treatment actually administered, $Y = y|T = t$. This is the *consistency assumption.*

Such untestable assumptions are not necessary in the framework proposed in this thesis. It is in fact the desire to avoid making such assumptions that is partly responsible for the development of the framework. It is difficult to understand whether assumption 3 is plausible or indeed what it means. Is there a potential response to each possible treatment floating around waiting to be revealed, or are there just a limited number specific to the context we are interested in? It is hard to imagine and impossible to verify.

In the counterfactual framework reviewed in section 3.3, a more intuitive argument is made for the existence of potential responses by presenting them as solutions to sets of structural equations that represent causal relationships.

Joint distributions of potential responses

The second problem is of a more practical nature and is termed the *fundamental problem of causal inference* by Holland (1986). From observed data in an experiment, it is only possible at best to estimate the marginal distributions of the potential responses. The observed data can tell us nothing about the correlation structure between them. In order to make inference about the joint distribution, it is necessary to make assumptions about the correlation

structure, or impose some constraints that will do this.

Such assumptions include *unit homogeneity* and *treatment unit additivity* amongst others. These assumptions are all discussed in section 3.5.1, as they are not unique to Rubin's model.

Another type of assumption is that of *Stable unit-treatment-value assumption* (SUTVA). This is made in order to exclude situations where there is interference between units. Again, this assumption is not unique to Rubin's model, and is discussed in section 3.5.1.

These assumptions are necessary given the initial assumption of the existence of potential responses and the definition of causal effects in terms of potential responses. Without making at least one such assumption, it would be impossible to evaluate causal effects and related parameters of interest such as variance of the effects etc. However, the assumptions are not testable, and making different assumptions can lead to different conclusions. (See Dawid 2000 for examples of the consequences of different correlation constraints on estimation, and *The problem with counterfactuals* in section 3.5.1.)

The relationship between decisions and causes

Rubin implicitly relates cause to decisions. He does this by stating that when relating the potential outcomes framework to the real world, it is important to only consider *a series of actions that could be applied to each experimental unit* (Rubin 1978). Hence he does not consider cases where for example the sex of a person is a cause of discrimination, because it is not possible to apply

a change of sex.

Notwithstanding Rubin's description of causes in terms of decisions, he never explicitly includes decisions as variables or conditions in his framework. In fact, he goes so far as to implicitly assume that a potential response is the same regardless of how the treatment was administered, in an experiment or under uncontrolled circumstances. That is, the difference between examples 2.3.1 and 2.3.2 is not taken into consideration. (If we further assume ignorability holds, which is generally the case. See assumption 2 in *Identification from observational data* in section 3.5.1.).

Assumption 4 Invariance of outcomes to treatment assignment mechanism
The potential response $Y^t(u)$ is the same regardless of the mechanism that assigns treatment to u.

From the point of view of the decision theoretic framework proposed here, this is a gross oversight. It is only when it is possible to derive expressions for probabilities conditional on interventions in terms of probabilities conditional on observation that we can make causal inference from observational data. Simply assuming that the outcomes are invariant to the administration is not sufficient and leads to false inference. Clearly, Rubin's inference is not to be dismissed, however, such assumptions can be avoided as we have seen in Chapter 2, and valid inference can be made.

3.2.4 Causal Inference as a Missing data problem

Rubin defines the causal effect of a treatment with respect to another in equation (3.1). Furthermore, he assumes that although only one potential response can become a realised outcome and be observed, both exist. Given these premises and Rubin's background in Bayesian analysis with missing data (Rubin and Little 2002), he tackles the problem of estimating causal effects as a missing data problem.

Potential responses, missing responses, observed responses

In order to give a brief overview of the Bayesian model used by Rubin, and discussed in section 3.2.5, some additional notation needs to be introduced. In particular the missing data notation.

Let $\mathbf{Y}$ denote all the responses that would have been observed if no data had been missing, that is, if it were possible to observe the response for both treatments $T = 1$ and $T = 0$. Let $\mathbf{Y} = (Y_{obs}, Y_{mis})$, where one of the two potential responses will be equal to Y_{obs} and the other to Y_{mis} as only one can be observed for every unit. The observed value of the observed response is denoted by y.

Similarly, let $\mathbf{X}$ represent the covariate data if all of it had been observed, then $\mathbf{X} = (X_{obs}, X_{mis})$. The observed value of the observed covariates is denoted by x.

We denote the probability density function of the observed and missing

variables[1] as follows;

$$f(Y_{obs}, Y_{mis}, X_{obs}, X_{mis}) \tag{3.3}$$

In the example, $Y_{obs} = (y^0(u_1), y^1(u_2), y^0(u_3), y^1(u_4), y^1(u_5))$,
$Y_{mis} = (y^1(u_1), y^0(u_2), y^1(u_3), y^0(u_4), y^0(u_5))$,
$X_{obs} = (x(u_1), x(u_2), x(u_4), x(u_5))$ and $X_{mis} = x(u_3)$.

Missing data indicators

To make inference about the joint distributions of observed and missing variables, it is necessary to include the mechanisms that control the missingness in the model. The missingness can be expressed in terms of variables called *missing data indicators.*

Let M be the missing data indicator for Y. In particular, let $M = (M^1, M^0)$, where $M^t = 0$ if Y^t is missing and 1 otherwise, for $t = 0, 1$. As only one of the potential response variables is ever observed, at least one of the missing data indicators will be 1 and the other 0 for each unit. It is possible for both of them to be missing if the data are lost or the patient does not report his headache status for example.

For the covariate variable, let M^X be the missing data indicator. For each patient, this takes on value 0 if the data are missing and 1 if they are observed.

Example 3.2.2 *Table 3.2 shows the same problem given in example 3.2.1*

[1]It should be conditional on a parameter π, but as it is not used for the purposes of this discussion, it is omitted for the sake of simplicity

with the addition on missing data indicators. As there are only two treatments and T takes on values $0, 1$, the column for the missing data indicator for treatment $T = 1$ is identical to that of treatment. If we were considering multiple treatments, then these would be different For example, u_3 did not

	Covariate		Treat	Response		Missing Data Ind.		
	A	S	T	Y^1	Y^0	M^X	M^1	M^0
u_1	0	0	0		1	1	0	1
u_2	1	0	1	0		1	1	0
u_3			0		0	0	0	1
u_4	2	1	1	1		1	1	0
u_5	1	1	1	1		1	1	0

Table 3.2: Table 3.1 with additional columns for missing data indicators.

have its covariates recorded, hence $M_3^X = 0$.

Note that the values of the missing data indicators are assumed to be observed always.

3.2.5 Bayesian probabilistic model

There are two aspects of the Bayesian approach in the Rubin causal model worth noting. The first is that by using the full Bayesian method, it is (at least in theory) possible to derive a predictive distribution for each of the potential responses, based on the priors (which are assigned for all the variables in the model) and the observed data. The second aspect is that if certain assumptions about the way the data is missing (or observed) are

made, then it becomes possible to estimate the predictive distributions from the observed data.

I will not describe the Bayesian methodology or discuss constraints on the prior distributions, as these are not relevant to the causal aspect of Rubin's model. For details on these aspects, see Rubin (1978). However, the assumptions necessary to make estimation of the predictive distributions possible, are of importance. First of all the assumption of exchangeability of study and future units is made.

Ignorability

The random variables in the aspirin example are (X, Y, T, M). Note that X Y and M represent the collection of potential reposes and potential missing data for every possible treatment under consideration in the domain of T. This is generally limited to two treatments, however, Rubin's theory caters to larger groups of treatments.

A possible factorisation of the joint distribution of the random variables is given in (3.4) below.

$$f(X,Y)k(T|X,Y)g(M|X,Y,T), \tag{3.4}$$

where $f(X,Y)$ is the marginal distribution of the potentially observable variables. In the running example, this is the marginal distribution of the covariates sex and age, and the response to treatment, that is, taking aspirin or not. $k(T|X,Y)$ is the probability of a treatment T given the (X,Y). That is, the probability of aspirin having been administered given the covariates and

responses. This term is the *assignment mechanism.* Finally, $g(M|X,Y,T)$ is the distribution of the recorded values in (X,Y) given the treatment mechanism. This term is the *recording mechanism.*

The aim of the above factorisation of the joint distribution, is to simplify the problem so that all inference can be based on what has been observed and not on the missing data. Data can be missing in two ways, either it is missing because it has been lost, or not recorded etc, or it is missing systematically, as in the case of potential responses.

Definition 3.2.1 (Ignorable Treatment Mechanism) *Let T be such that only two treatments are possible. A treatment mechanism is ignorable if*

$$T \perp\!\!\!\perp Y_{mis}, X_{mis} | Y_{obs} = y, X_{obs} = x. \tag{3.5}$$

for all patterns of Y_{mis} and X_{mis}.

Then, the treatment assignment mechanism can be expressed solely in terms of the observed data;

$$k(T|X,Y) = k(T|X_{obs}, Y_{obs}). \tag{3.6}$$

Similarly,

Definition 3.2.2 (Ignorable Recording Mechanism) *A recording mechanism is ignorable if*

$$M \perp\!\!\!\perp Y_{mis}, X_{mis} | Y_{obs} = y, X_{obs} = x, T. \tag{3.7}$$

for all Y_{mis} and X_{mis}.

As with the treatment mechanism, the recording mechanism becomes

$$g(M|X,Y,T) = g(M|X_{obs},Y_{obs},T). \tag{3.8}$$

Consider example 2.3.1. Are the treatment and recording mechanisms ignorable? In order to establish this, we need to know more about the experimental setup and make some additional assumptions.

Say that the treatments were randomised, that all the patients were given the same dose of aspirin or placebo and that they all took their assigned treatments and truthfully reported their responses to treatment.

Further, assume that the missing covariate information is missing at random (see definition 3.2.3 below or (Rubin and Little 2002) section 1.3). We see that none of the realised outcome data is missing.

Given this information, we can divide the recording mechanism further into two parts $g_1(M^0, M^1|X,Y,T,M^X)$ and $g_2(M^X|X,Y,T)$. g_1 represents the missingness of the data, and g_2 represents the missingness of the covariate information.

Now, as the treatment assignment mechanism is random, it is in fact ignorable. As the covariate information is missing at random, g_2 is also ignorable. Finally g_1 is fully known, as it depends entirely on the treatment assignment mechanism, and is therefore also ignorable. Hence, given the assumptions about the data recording and reporting, and the treatment assignment, the example has both ignorable treatment and recording mechanisms. For a detailed discussion of the conditions necessary for ignorability see Rubin (1978).

Comments on *missingness*

From the idea of missingness comes the idea of *missing at random* (Rubin and Little 2002 section 1.3).

Definition 3.2.3 *Missing at Random*

Missing data for a variable Y is said to be ***missing at random*** *if the probability of missing data on Y is unrelated to the value of Y, after controlling for the variables in the analysis.*

Thus in the example used above it may be plausible to assume that the probability that the covariate information is missing is unrelated to the value of the covariates, having taken into consideration the other variables in the problem. Now this assumption is clearly not testable because we will never know the age and sex of those who do not report it. If however, we believed that the sample of patients participating in the clinical trial has a large component of ladies who prefer not to reveal their age, then this assumption would not be plausible.

There are some objections to using missing data methods for causal inference. The first is that covariate information is in principle observable, whereas the unrealised outcome is never observable. Using methods developed to tackle the missingness of the former on the latter requires some strong untestable assumptions, which are not justifiable in my opinion. The second objection is that the missing data notation does not cater for the relationship between interventions and causes. Even though Rubin states that these are

related, he does not introduce this explicitly.

The concept of *missingness* is not to everyone's liking. Consider the individual causal effect in (3.1). Say the individual has been treated, and we have observed y^1, and hence not observed Y^0. Rubin only considers that Y^0 is missing, however, $y^1 - Y^0$ as well as any other function of Y^0 is also missing. Further, Rubin constructs much of his theory considering properties of missingness. These properties are generally assumed as it is not possible to validate them, as the data is missing.

It is possible to reconsider the concept of missingness in terms of what is observable, as seen in Dawid and Dickey (1977).

3.3 Functional - Graphical frameworks

The following section covers causal frameworks that use counterfactuals and graphical models. I focus on the framework proposed by Pearl (2000) as I feel that it sufficiently represents equivalent frameworks such as those put forward by Spirtes et al. (2000) (although these are purely graphical), Robins (1986), and in the Econometrics literature by Heckman in particular starting in Heckman and Robb (1985). Robins' application of this type of framework to time-varying treatment regimes and the *G-computation* formula are noteworthy and are discussed in some detail in section 3.3.6.

The fundamental concept in Pearl's counterfactual framework are *functional causal models* (FCMs), which in turn are based on the *structural equation models* (SEMs) popular in the social sciences. FCMs relate variables to

each other through deterministic functions.

The motivation for this approach is that most scientific research (excluding Quantum theory) is expressed in terms deterministic functions and that these are how the human mind intuitively understands causality. An example of this is the equation relating voltage to resistance and current given in example 2.1.2. The assumption underlying the standard use of this formula is that if we measure these three variables in a circuit and find a deviation from the formula, then it is due to measurement error.

However, as it is impossible to avoid random error, Pearl introduces *Probabilistic causal models* (PCMs) by adding a stochastic element to FCMs. This addition is necessary to account for the inherent uncertainty in measurement and the lack of complete knowledge, and is in stark contrast to the decision theoretic approach proposed in this thesis, where probabilities are considered the most appropriate description of our perception of the world.

It is interesting that Pearl initially (Pearl 1993) introduced the intervention variable F_X. However, he found that just using the intervention variable and standard probabilistic concepts did not answer all the causal questions he was interested in. In particular those of the type *would my headache have gone away if I had taken an aspirin?* needed to compare what happened to what did not happen in the past, could not be tackled without counterfactuals. These questions are questions that are not considered useful in the causal framework proposed in this thesis as argued in section 2.2.4, and useful questions can be formulated in their stead. In this case, *will my headache go*

away if I take an aspirin?

In this review of his contribution to causal inference, attention will be restricted to the deterministic and counterfactual aspects of Pearl's framework.

The basic building blocks of the functional model of causality are structural equation models, counterfactual notation and graphs. The graphical aspect is used to describe causal structures and to determine whether quantities of interest are identifiable or not. The SEMs are used to describe the individual relationships between an effect and its causes, and they reflect the parent-child relationships in the corresponding graph. Finally, counterfactuals are solutions to the structural equations for values of the causes that did not come to be. The graphical aspects will be covered in the section 3.3.1, the SEMS in section 3.3.2 counterfactuals will be tackled in section 3.3.4. The problem of identification is looked at in section 3.3.5. The section ends in an example of how to estimate a causal effect using counterfactual notation and finally some important assumptions underlying the counterfactual approach are discussed in section 3.3.5.

3.3.1 Graphs

The graphical notation used by Pearl is analogous to that defined in section 2.4. Similarly, he adopts the *Markov condition* also defined in this section, and generally limits his attention to causal structures that can be described by DAGs.

One of the most important graphical concepts in this framework is the

graphical equivalent of conditional independence. This concept is central as it determines questions of identification (section 3.3.5) and is thus essential to the estimation of causal effects.

Definition 3.3.1 (d-separation) *A path p is said to be d-separated by a set of nodes Z if and only if*

1. *p contains a chain $i \rightarrow m \rightarrow j$ or a fork $i \leftarrow m \rightarrow j$ such that $m \in Z$, or*

2. *p contains an inverted fork $i \rightarrow m \leftarrow j$ such that $m \notin Z$ and such that no descendant of m is in Z.*

A set Z d-separates X from Y if and only if Z d-separates every path from X to Y. We denote this by $X \perp_d Y|Z$.

When a DAG corresponds to a probability distribution, the d-separation criterion leads to the same results as the moralisation criterion, with the d-separation symbol $\perp_d$ replacing the conditional independence symbol $\perp\!\!\!\perp$. For example, if the DAG in figure 3.1 represents the conditional independence structure of the joint probability distribution of X, Y,Z and W, then we can use the moralisation criterion to determine that $X \perp\!\!\!\perp W|Z$. Similarly, we can use the d-separation criterion to determine that Y d-separates X and Z. The two criteria are in fact equivalent when the DAG represents the conditional independence structure of the joint probability distribution of the variables. This is formalised in Theorem 1.2.4 in Pearl (2000).

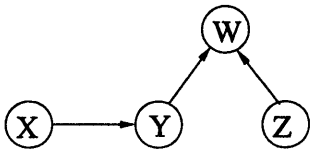

Figure 3.1: Y d-separates W and X, W does not d-separate Y and Z.

3.3.2 Structural Equation Models

Structural equation models were first used in economics and the social sciences, where they are still popular in the literature now. Relationships between the observed variables are assumed to be causal and known. These are then be expressed by means of a set of functions, describing the quantitative aspect of the relationships. A graph that describes the relationships between effects and causes as parent child relationships can then be drawn. Consider the following example;

Example 3.3.1 *Congestion Charge*

Almost three years ago the *congestion charge* was introduced in central London. Any private vehicle entering a delimited central area is charged 5 pounds sterling. The Mayor of London commissioned a study to see how effect of the congestion charge on pollution via traffic to determine whether to extend the area. The assumed causal structure between the three variables and their errors is given in figure 3.2. C is an indicator representing the congestion charge, T the traffic levels and Y the pollution levels in the city. The errors are further assumed to be independent of one another.

There are occasions in which this assumption cannot be made. For in-

stance if we believed that the traffic and the pollution were both affected by an unobserved variable, then we would replace U_3 with U_2.

The equations that correspond to this graph are given below:

$$\begin{aligned} C &= f_1(U_1), \\ T &= f_2(C, U_2), \\ Y &= f_3(T, U_3). \end{aligned}$$

If there is evidence to do so, a more specific functional form can be imposed. For example, we may have reason to believe that the pollution is related linearly to the traffic and we can then replace $Y = f_3(T, U_3)$ with $Y = \alpha + \beta T + U_3$.

The SEMs can be used as a tool for inference (formalised in definition 3.3.2) as follows: *1)* If we *set* $T = t_0$ then we replace $T = f_2(C, U_1)$ with $T = t_0$ as there is no uncertainty remaining about the value of T, and also *2)* we insert t_0 into the equation for Y; $Y_{t_0} = f_3(t_0, U_3)$, while the equation for C remains unaffected. Thus if we assume the linear relationship $Y = \alpha + \beta Y + U_3$, after intervention on T, we have that $Y = \alpha + \beta t_0 + U_3$.

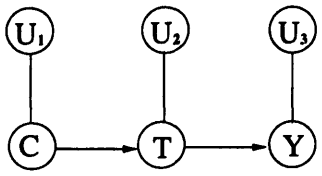

Figure 3.2: The assumed relationships between C congestion charge, T traffic levels, Y pollution levels.

3.3.3 The *do* operator

SEMs and their corresponding graphs can be used to predict the effects of interventions, which Pearl defines as *causal effects* by introducing the equivalent to the intervention node F_X, the $do()$ operator

The *do* operator essentially sets the value of a variable to a particular value. In fact if X is a random variable, then $do(X = x)$ is identical to $F_X = x$ where F_X is the intervention node on X and x is a realisation of X.

Look at the congestion charge example above. Saying $do(C = c)$ means that we are intervening on the congestion charge and setting it to c. If $C = 1$ means that a city is introducing the congestion charge, then the *causal effect* of introducing the congestion charge on traffic is $p(T = t|do(C = 1))$, which gives the probability of $T = t$ induced by deleting $C = f_1(u_1)$ and replacing $T = f_2(C, U_2)$ with $t_1 = f_2(1, U_2)$.

3.3.4 Counterfactuals and the Causal Model

In this section we look at Pearl's *functional causal model* (see Definition 7.1.1 in Pearl (2000)). This is made up of the observed variables in the problem, the error terms associated with these variables and the set of equations that relates the causes and their effects. The definition is given below as it is essential for the understanding of the functional model.

Definition 3.3.2 *A **Causal Model** is a triple* $\mathbf{M} =< \mathbf{U}, \mathbf{V}, F >$ *such that*

1. $\mathbf{U}$ *is a set of background/error variables that are determined by factors*

outside the model;

2. $\mathbf{V}$ *is the set of variables* $\{V_1, \ldots, V_n\}$, *functionally determined by* $\mathbf{V} \cup \mathbf{U}$. *The variables we are interested in including in the model;*

3. F, *the set of functions* $\{f_1, \ldots f_n\}$ *such that each* f_i *is a mapping from* $\mathbf{U} \cup ((V) \setminus V_i)$ *to* V_i *and such that the entire set* F *forms a mapping from* $\mathbf{U}$ *to* $\mathbf{V}$. *Thus,* f_i *determines the value of* V_i *given the values of all other variables in* $\mathbf{U} \cup \mathbf{V}$. *There is a unique solution for each function for* F *given by* $V(u)$;

A graph given by $G(\mathbf{M})$ is associated to the causal model $\mathbf{M}$ and is such that each variable corresponds to a node and the directed edges link the effects as children and the causes as parents.

Let X be a variable in $\mathbf{V}$ and x_0 a realisation of X. A submodel M_{x_0} can be derived from M by replacing the function $x = f_i(pa_i, u_i)$ with $x = x_0$ and replacing instances of x with x_0 in the remaining functions. This can be extended to a subset $\mathbf{X}$ of $\mathbf{V}$, inducing submodel $M_{\mathbf{x}}$, where $\mathbf{x}$ is the vector of realisations of $\mathbf{X}$ and where $F_{\mathbf{x}} = \{f_i : V_i \notin \mathbf{X}\} \cup \{\mathbf{X} = \mathbf{x}\}\}$ is the set of equations induced by setting $\mathbf{X}$ to $\mathbf{x}$. The *effect of an action* $do(X = x)$ is the submodel M_x.

Note that it is not necessary to set a variable $do(X = x)$ in order to be able to estimate the causal effect of this setting, as we can solve M_x for a particular response of interest. Thus we see that the functional equation approach is conducive to the counterfactual approach. Looking at the congestion charge

example, if pollution can be expressed as a function of traffic and an error, and the function can be estimated, then it is not necessary for the traffic to ever attain a value t to be able to calculate the resulting levels of pollution.

This leads to the following definitions for potential responses and counterfactuals:

Definition 3.3.3 *Potential responses and counterfactuals*
Let $\mathbf{X}$ *and* $\mathbf{Y}$ *be two subsets of variables in* $\mathbf{V}$. *The* ***potential response*** *of* Y *to action* $do(\mathbf{X} = \mathbf{x})$ *on a particular unit* u *is denoted* $\mathbf{Y_x}(u)$, *and is the solution for* $\mathbf{Y}$ *of the set of equations* $F_{\mathbf{x}}$. *The* ***counterfactual*** *"the value that* $\mathbf{Y}$ *would have taken had* $\mathbf{X}$ *been* $\mathbf{x}$*" is* $\mathbf{Y_x}(u)$.

There is no difference between a potential response and a counterfactual in terms of the mathematics, it is simply a matter of timing as a potential response refers to the solutions to all the possible actions, whereas the counterfactual is the solution to a specific unrealised action. Note also that u, the unit (which can be an individual or a specific set of characteristics) is itself also unchanged by intervention.

Pearl introduces the stochastic element to account for incomplete knowledge and measurement error:

Definition 3.3.4 *A* ***Probabilistic causal model*** *is a pair* $< \mathbf{M}, P(u) >$ *where* $\mathbf{M}$ *is a causal model, and* $P(u)$ *is the probability function defined over the domain of* $\mathbf{U}$.

For any variable Y in $\mathbf{V}$, the probability $p(Y = y)$ is defined as a function of the background variables as follows

$$p(Y = y) := \sum_{\{u|Y(u)=y\}} P(u).$$

This can be extended to counterfactual probabilities;

$$p(Y_x = y) := \sum_{\{u|Y_x(u)=y\}} P(u).$$

As mentioned in the introduction to the functional framework, probabilities are not the basic building block, rather, the functions are taken as basic, and the probabilistic element is added on to account for the randomness that ensues from our incomplete knowledge of the causal structures at work. As it is impossible to avoid the probabilistic element, the causal effect is defined as a probability;

Definition 3.3.5 *Causal Effect*
Given a causal model $\mathbf{M}$*, and two disjoint sets* $\mathbf{X}$ *and* $\mathbf{Y}$ *in* $\mathbf{V}$ *, the* ***causal effect*** *of* $\mathbf{X}$ *on* $\mathbf{Y}$ *is given by* $p(Y = y|M_x)$ *for the realisation* $\mathbf{x}$ *of* $\mathbf{X}$.

Note that the causal effect is not defined comparatively here, however, it is generally a comparison of such causal effects that is of interest.

3.3.5 Identifiability

As discussed in section 2.3.3, the data that is available to the statistician is often observational. In order to make inference about causal effects it is therefore necessary to make additional assumptions or explore the conditional

independence / d-separation relationships between the observed variables. Pearl's approach to identification is based on the manipulation of graphical models and counterfactual conditional independences.

The process of using graphs to identify a causal quantity from observational data is analogous to the non-counterfactual use of conditional independences; the graph is used to determine the d-separations between the cause variable and the effect variable, if the set of observed variables that d-separate the cause and the effect variable obey a set of constraints then the causal effect is identifiable from observational data.

Back-door criterion

The simplest constraint is the *back-door criterion* (Pearl 2000 definition 3.3.1) given below. It determines whether a subset of observed variables that d-separates the causal variable from the effect variable are sufficient to allow the identification of the causal effect from observational data. In particular, it is useful when there are no unobserved confounders.

A set of variables Z satisfies the *back-door criterion* relative to an ordered pair of variables X, Y in a DAG G if:

1. no node in Z is a descendant of X; and

2. Z blocks every path between X and Y that contains an arrow into X.

As a consequence, the causal effect of X on Y is given by

$$p(Y_x = y) = p(Y = y|do(X = x)) \quad =$$

$$\sum_z p(Y = y|Z = z, X = x)p(Z = z). \quad (3.9)$$

A similar result can be derived in terms of the augmented DAG notation. If we take the above variables, then $do(X = x)$ is equivalent to $F_X = x$, so we have that

$$p(Y = y|F_X = x) = \sum_z p(Y = y|Z = z, X = x, F_x = \emptyset)p(Z = z|F_X = \emptyset)$$

if

1. $Z \perp\!\!\!\perp F_X$ and

2. $Y \perp\!\!\!\perp F_X|X, Z$.

do calculus

The back-door criterion is an example of the use of *do calculus*, a set of graphical rules intended to enable identification of experimental quantities using observational data. To show the rules, it is first of all necessary to introduce some additional notation. Let X, Y and Z be disjoint nodes in G a DAG with associated probability distribution $p()$. Let $G_{\overline{X}}$ denote G excluding the edges pointing into X. Also, let $G_{\underline{X}}$ denote G excluding edges pointing out of X. Combining the notation we denote by $G_{\overline{X}\underline{Z}}$, G excluding edges pointing into X and edges pointing out of Z.

3.3.5 Rules of do calculus

1. insertion/deletion of observations

   $p(Y|do(X = x), Z, W) = p(Y|do(X = x), W)$

   if $(Y \perp_d Z|X, W)_{G_{\overline{X}}}$

2. action/observation exchange

   $p(Y|do(X=x),do(Z=z),W) = p(Y|do(X=x),Z,W)$

   if $(Y \perp_d Z|X,W)_{G_{\overline{X}\underline{Z}}}$

3. insertion/deletion of actions

   $p(Y|do(X=x),do(Z=z),W) = p(Y|do(X=x),W)$

   if $(Y \perp_d Z|X,W)_{G_{\overline{X}_{\underline{Z(W)}}}}$.

   Where $Z(W)$ is the set of nodes in Z that are not ancestors of any nodes in W in $G_{\overline{X}}$.

These rules can be used to infer an exhaustive list of constraints for graphical identification of causal quantities from observational data. These are given in Pearl (2000) 3.5, however, a detailed description of these is beyond the scope of this review.

As in the case of the back-door criterion, these graphical rules can be translated into the augmented notation. To do this, consider the following. Say we have a generic conditional independence $A \perp\!\!\!\perp B|X,C$ where A, B, and C are nodes or sets of nodes in the graph G of interest. Now, when $F_X = x$ then X is identical to F_X. Thus any conditional independence involving X extends to F_X. So $A \perp\!\!\!\perp B|F_X = x, X = x, C$. If $F_X = \emptyset$ then it is trivially true that $A \perp\!\!\!\perp B|F_X = \emptyset, X, C$. Thus, if $A \perp\!\!\!\perp B|X,C$ then $A \perp\!\!\!\perp B|F_X, X, C$.

When we look at the rules of do calculus from the decision theoretic point of view, we see that the conditional independences can be rewritten using F_X.

3.3.5 Rules of do calculus in decision theoretic framework

1. $p(Y|F_X = x, Z, W) = p(Y|F_X = x, W)$ if $Y \perp\!\!\!\perp Z|X, W$.

   Now $p(Y|F_X = x, Z, W) \equiv p(Y|F_X = x, X = x, Z, W)$ as F_X and X are identical when $F_X = x$. Also, the conditional independence extends to $Y \perp\!\!\!\perp Z|X, F_X, W$. So $p(Y|F_X = x, X = x, Z, W) = p(Y|F_X = x, X = x, W)$. Finally, again by fact that F_X and X are identical when $F_X = x$, $p(Y|F_X = x, X = x, W) = p(Y|F_X = x, W)$.

2. $p(Y|F_X = x, F_Z = z, W) = p(Y|F_X = x, Z, W)$ if $Y \perp\!\!\!\perp F_Z|X, Z, W$ by a similar argument as the one above, similarly

3. $p(Y|F_X = x, F_Z = z, W) = p(Y|F_X = x, W)$ if $Y \perp\!\!\!\perp Z|X, W$.

Note that the first and last conditions are the same condition in the decision theoretic framework.

Note also that when a quantity is not identifiable from the graph, it is possible to make additional assumptions that make the quantity identifiable. These assumptions tend to involve imposing specific functional forms on the relationships between the variables. Alternatively *separability* assumptions can be imposed. An example of this is given in chapter 5 section 6.6.1.

Calculating a causal effect using counterfactuals

Say we have a causal model **M** with the components given in definition 3.3.2 with a non-parametric set of structural equations, and a corresponding graph G. To evaluate the causal effect of one variable on another, we first find the set of conditional independences that hold between the potential response (and

hence counterfactual) variables. These are then used to evaluate the causal effect. Note that notwithstanding the name *restriction* given to following statements,, they are not additional restrictions but consequences of the rules of do calculus and graphical manipulation.

Exclusion restriction:

For every variable Y in $\mathbf{V} \cup \mathbf{U}$ with parents PA_Y and for every set $\mathbf{Z} \in \mathbf{V}$ such that $\mathbf{Z} \cap \mathbf{PA_Y} = \emptyset$ we have that

$$Y_{\mathbf{pa}_Y}(u) = Y_{\mathbf{pa}_Y \mathbf{z}}(u), \tag{3.10}$$

where $Y_{\mathbf{pa}_Y \mathbf{z}}(u)$ is the potential response to any setting of the parents of Y and any setting of the set of variables $\mathbf{Z}$.

Independence restriction:

If $Z_1, \ldots, Z_k$ is a set of variables in $\mathbf{V}$ not connected to Y via paths containing only $\mathbf{U}$ variables in the graph G, then

$$Y_{\mathbf{pa}_Y} \perp\!\!\!\perp \{Z_1 \mathbf{pa}_{Z_1}, \ldots, Z_k \mathbf{pa}_{Z_k}\}. \tag{3.11}$$

The exclusion restriction condition essentially states that if $\mathbf{Z}$ are variables disjoint from the parents of Y then Y is independent of whether Z has been set or what it has been set to conditional on its parents having been set. For example, if the $\mathbf{Z}$s are descendants of Y in the graph, then it is easy to see that Y does not change if $\mathbf{Z}$ are set given that the parents of Y have been set. That it $Y \perp\!\!\!\perp \mathbf{F_Z} | \mathbf{F_{PA_Y}}$.

The independence restriction condition says that if two variables are not connected by paths that only contain unobserved background/error variables then they are marginally independent given their parents have been set. How the two restrictions work is best shown in an example.

Consider example 3.3.1. The problem is very simple as we assume that the Us are all mutually independent of one another. The problem would be complicated if we assumed an unobserved common parent for Y and T for example, as this would violate an independence restriction.

First by exclusion restriction we have that

ER1 $T_c(u) = T_{cp}(u)$,

ER2 $C_y(u) = C_{yt}(u) = C_t(u) = C(u)$,

ER3 $Y_t(u) = Y_{tc}(u)$.

ER4 By construction we have that $Y_c(u) = Y_{t_{c(u)}}$ if $T_c(u) = t$. By ER3 it follows that $Y_c(u) = Y_{c_{t_c}}(u) = Y_t(u)$ when $T_c(u) = t_c$.

Then, by independence restriction

IR1 $T_c \perp\!\!\!\perp (C, Y_t)$,

IR2 $Y_t \perp\!\!\!\perp C$.

The process of calculating $p(Y_c)$ given the above rules is as follows;

$$p(Y_c) = p(Y_{t_c} = y) \qquad \text{by ER4,}$$

$$
\begin{aligned}
&= \sum_t p(Y_{t_c} = y | T_c = t) p(T_c = t) \\
&= \sum_t p(Y_t = y | T_c = t) p(T_c = t) && \text{by ER3} \\
&= \sum_t p(Y_t = y) p(T_c = t) && \text{by IR1.} \quad (3.12)
\end{aligned}
$$

Next we evaluate

$$
\begin{aligned}
p(T_c = t) &= p(T_c = t | C = c) && \text{by IR1} \\
&= p(T = t | C = c),
\end{aligned}
$$

and similarly

$$
\begin{aligned}
p(Y_t = y) &= p(Y_t = y | T = t) \\
&= p(Y = y | T = t).
\end{aligned}
$$

These we can then substitute into (3.12) and calculate $p(Y_c)$ the causal effect of having a headache:

$$
P(Y_c) = \sum_t p(Y = y | T = t) p(T = t | C = c)
$$

In this simple example, the structural equations are not specified and are not necessary. They are used when the relationship between the variables is either known or needs to be assumed in order to make inference.

If the question we want to ask is, *if I introduce the congestion charge, will pollution levels change?* then this is a problem that can be solved more simply in decision theoretic terms; The DAG associated to the problem is the DAG in figure 3.3 without the Us as we assume there are no dependences between the observed variables other than the ones depicted. C, represents

the introduction of the congestion charge, and it is not easy to see how it could be a chance variable. Thus let us assume that F_C cannot be idle. The

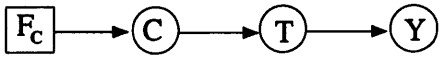

Figure 3.3: Congestion Charge example in non-counterfactual terms

causal quantity of interest is expressed in the non-counterfactual framework as $p(Y = y|F_C = c)$, the probability of pollution level y given the congestion charge is set at c.

$$
\begin{aligned}
p(Y = y|F_C = c) &= \sum_t p(Y = y|T = t, C = c, F_C = c)p(T = t|C = c, F_C = c) \\
&\quad \text{as } p(C = c|F_C = c) = 1 \text{ so no need to sum over } C \\
&= \sum_t p(Y = y|T = t, F_C = \emptyset)p(T = t|C = c, F_C = c) \\
&\quad \text{as } Y \perp\!\!\!\perp (C, F_C)|T. \\
&= \sum_t p(Y = y|T = t, F_C = \emptyset)p(T = t|C = c, F_C = \emptyset) \\
&\quad \text{as } T \perp\!\!\!\perp F_C|C.
\end{aligned}
$$

If we have two settings of the congestion charge, 1 and 0 for introducing it and not introducing it respectively, then we can estimate $p(Y = y|F_C = c)$ for both these settings, and the causal effect would be the difference of the expectations.

If we want to answer the question *would the pollution levels have changed if we had not introduced the congestion charge?* then it is necessary to use the counterfactual notation. Note that the quantity of interest remains the same, that is $Y_c(u)$. It is also the quantity of interest for of the type, *the pollution*

levels have changed, is it because we introduced the congestion charge?. Yet these questions are of a completely different nature, should there not be a different method to tackle each one?

Assumptions underlying the use of counterfactuals

Pearl's functional framework rests on the assumption that causal relationships can be adequately described by deterministic functions. From this fundamental assumption, the rest follows.

Assumption 5 Causal relationships are deterministic functions, and due to incomplete knowledge, it is appropriate to model causal relationships as deterministic functions with an error term.

This also begs the question, *is there a limit to the number of possible actions?* Can we in theory determine the consequences of actions that cannot take place if we define a set of structural equations in a particular manner?

It follows from adopting the deterministic view point that the structural equations are invariant to intervention. Each equation is autonomous and describes the individual relationship between a variable and its causes of interest. Although they can be solved for a particular intervention, the intervention only affects the variable that has been intervened upon and its descendants, leaving the remaining variables unaffected. Thus, we can solve for an intervention that did not take place, simply by setting an equation to a constant and solving the related functions. Thus, it is not necessary to intervene to be able to evaluate the consequences of such an intervention.

Further assumptions are made which are common to Rubin's potential outcomes framework. However, it is worth noting that where Rubin assumes the existence of potential responses, their existence follows from Pearl's assumption that a particular data situation can be described by deterministic functions.

It is also worth reiterating that the counterfactual (Y_x) notation hides the *do* notation in subscripts, and as we see in the example in the section 3.3.5, the subscripts are used even when an intervention has not taken place. This also happens in the decision theoretic framework as $p(Y|F_X = x, X = x) \equiv p(Y|X = x, F_X = \emptyset)$ when $Y \perp\!\!\!\perp F_X|X$. However, whilst the notation in the counterfactual framework obscures the links between actions and effects, the notation in the decision theoretic framework shows the link explicitly.

3.3.6 Time varying treatments and G-computation

This review of Robin's work in this section is based on Lok (2001), which in turn is based on Robins (1995) and (1998). The basic problem in discrete time is described and the relevant notation introduced. Finally the *G-computation formula* for estimation of causal effects from observational data under the *no unmeasured confounders assumption* is stated and interpreted.

Robins' approach, has elements of both the functional model (he uses extensions of SEMs called *structural nested failure time models* amongst others) and the potential outcome model (he makes assumptions such as *consistency*). His interest lies in identifying causal effects of time varying treatments

from observational data where only some treatment regimes have been administered and the process is not controlled in the sense that it is not an experimental setting. This is due to the sources of data which are records of treatments on patients suffering from long-term, potentially fatal diseases such as HIV/AIDS or cancer. Thus treatments cannot be randomised, and the doctors must assess on the basis of past evidence what treatment to administer next.

The quantity of interest is the causal effect of treatment on survival time. Due to the nature of the diseases, no clinical trials to determine their efficacy can be run, and so any causal inference must be made from the observational data by making the appropriate assumptions.

The basic set-up is as follows; a patient has a set of characteristics, some that do not vary over time, such as sex, and others that do, such as variables that indicate the status of the disease. In the case of HIV/AIDS, this may be viral count. The patient is visited by a doctor over time and at every visit, the doctor must decide, based on the status of the patient, what treatment to prescribe. The collection of treatments over time is called a treatment *regime*. The treatment will depend on the time-varying patient characteristics and previous treatments received. The quantities of interest are the effects of the treatment regime received on the survival time. As there is no control group for these patients, and the approach adopted by Robins is counterfactual in nature, the effect is estimated by comparing the treatment regime that *was* received to a counterfactual treatment regime.

Notation

Consider the notation for a single patient. The time points at which a patient is visited are $\tau_0, \ldots, \tau_K$, such that $\tau_i = t$ means the patient's i^{th} visit took place at time t after his first visit. The patient characteristics at time τ_k are denoted by L_k for $k = 0, \ldots, K$, with realisation l_k and the treatment assigned by the doctor is denoted by A_k for $k = 1, \ldots, K$. The outcome of interest is survival time and is denoted by $Y \equiv L_K$. Finally the collection of characteristics up to visit k are denoted by $\overline{L}_k$ and treatments up to visit k are denoted by $\overline{A}_k$. This notation is necessary as treatments at time t are likely to depend on past treatments and covariate information.

Robins adopts a deterministic viewpoint and assumes that the patient characteristics and the treatments are related by a collection of functions. These are in fact the *regimes*. They are denoted by g^k with

$$g^k : \overline{\mathcal{L}}_k \rightarrow \overline{\mathcal{A}}_k, \tag{3.13}$$

where $\overline{\mathcal{L}}_k$ and $\overline{\mathcal{A}_k}$ are the domains of $\overline{L}_k$ and $\overline{A}_k$ for $k = 0, \ldots, K$. Thus the gs are the potential realisations of the As.

Figure 3.4 shows the general structure of Robins problem.

Counterfactual assumptions

In order to be able to evaluate the effect of a realised and thus observed treatment regime by comparing it to a counterfactual treatment regime, Robins must assume that for every possible treatment regime, there is a corresponding counterfactual outcome.

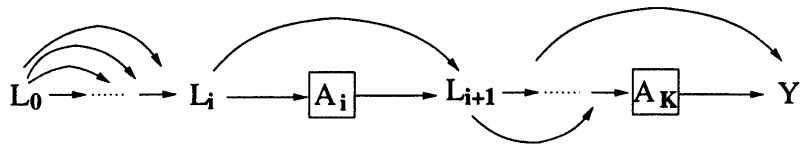

Figure 3.4: Robins problem of time-varying treatments in a graph. Each L variable has an arrow coming out of it going to every other L and every A has arrows coming from L's preceeding it. Only the first L shows this to avoid confusion.

Assumption 6 Existence of counterfactual random variables:

For every patient, there exists a random variable Y^g, the survival time had the patient received treatment regime g.

An additional counterfactual assumption is also made namely, that the solution to the regime that was applied to a patient must be the observed outcome. By Robin's definition (see (3.13)) g, the regime is a function, and thus can be solved for a particular set of past covariates and actions.

Assumption 7 Consistency:

For any fixed identifiable treatment regime g, $\bar{l}_k \in \overline{L}_k$ and $t \in (\tau_k, \tau_{k+1})$

$$\begin{aligned} &\{Y^g > t, \overline{L}_k = \bar{l}_k, \overline{A}_k = \overline{g}(\bar{l}_k), Y \geq \tau_k\} \\ &= \{Y > t, \overline{L}_k = \bar{l}_k, \overline{A}_k = \overline{g}(\bar{l}_k), Y \geq \tau_k\} \end{aligned} \tag{3.14}$$

This means that if the outcome of interest is survival time, then the patient would die under regime g if and only if he actually did die and received the same treatments as under regime g. Note that Y^g is a potential response (like Rubin conceives of them) and is equal to the realised response if the regime

g is the one actually performed. This type of assumption is also made by Rubin.

No unmeasured confounders

One problem with causal inference from non-experimental data, is that treatments are not generally set according to a predetermined treatment strategy, but by a doctor who is himself a part of the study. Hence they can depend on the outcome of interest through an unobserved confounder, such as the Doctor's treatment decision criteria. For example, if a doctor treats only patient who he thinks will react positively, and this is not taken into account when making inference, then the results will be incorrect. To overcome this problem, Robins assumes that there are no unmeasured confounders, that is, either the doctor clearly states his treatment selection criteria, or there are no unknown selection criteria, and the assignment is *randomised* given the past.

Assumption 8 No unmeasured confounders:
For any fixed treatment regime g, for any time τ_k and for any $\bar{l}_k \in \overline{\mathcal{L}}_k$ we have that

$$A_k \perp\!\!\!\perp Y^g | \overline{L}_k = \bar{l}_k, \overline{A}_{k-1} = \overline{g}(\bar{l}_{k-1}). \tag{3.15}$$

That is, given the past treatment regime and the patient characteristics up until time τ_k (which are common for both the counterfactual Y^g and the actual response Y i.e. $\overline{A}_{k-1} = \overline{g}(\bar{l}_{k-1})$), the current treatment does not depend on

what the outcome would be under regime g. It can also be interpreted as meaning that given past treatment and patient status, the current treatment is *random*. For a single treatment, this assumption is analogous to Rubin's conditional ignorability assumption.

G-computation

For discrete time points and for a single patient, under identifiable treatment regimes (that is, regimes that are actually possible, as every combination of characteristics and treatments is not necessarily possible) and the assumptions of consistency and no unmeasured confounders we have that

$$\begin{aligned} p(Y^g > t) \quad &= \quad \sum_{\bar{l}_0} \cdots \sum_{\bar{l}_j} [p(Y > t | \overline{L}_j = \bar{l}_j, \overline{A}_j = \overline{g}(\bar{l}_j), Y > \tau_j \\ &\prod_{m=0}^{j} \quad \{p(\overline{L}_m = \bar{l}_m | \overline{L}_{m-1} = \bar{l}_{m-1}, \overline{A}_{m-1} = \overline{g}(\bar{l}_{m-1}), Y > \tau_m) \\ &\times \quad p(Y > \tau_m | \overline{L}_{m-1} = \bar{l}_{m-1}, \overline{A}_{m-1} = \overline{g}(\bar{l}_{m-1}), Y > \tau_{m-1})\}]. \end{aligned} \tag{3.16}$$

This formula is arrived at by recursive application of (3.15). It has analogues in most other frameworks. If we consider a single treatment, with the associated DAG similar to the one in figure 3.4, we arrive at the back-door formula given in Pearl, (3.9). Also, it can be reformulated in non-counterfactual terms as in Dawid (2002) for the single treatment case. An extension with time varying treatments is being developed by Dawid and Didelez (personal communication).

3.4 Heckerman and Shachter's cause in terms of unresponsiveness

Heckerman and Shachter's (for-with HS) approach to causality is a generalisation of Pearl's framework with decision theoretic elements. It is most notable for the use of *unresponsiveness* as the basic concept in terms of which cause is defined and for explicitly allowing variables that cannot be intervened upon to be causes.

This section is based on Heckerman and Shachter (1995) and covers the basic set-up of the framework in terms of the decision maker, and goes on to define unresponsiveness and cause. Then it looks briefly at graphs in *canonical form*, which enable counterfactual to be explicitly represented in graphs. These are reminiscent of *twin networks* in Pearl (2000) 7.1.4.

3.4.1 Decisions

HS begin by defining a decision maker who can make one or more decisions. The decision maker exists in a world which has a number of possible *states of nature*. Every combination of an *act* (decision) with a state of nature, results in a deterministic mapping from these to a *consequence*. This set-up is based on Savage's (Savage 1954) decision theoretic primitives.

There are two types of variables in every problem, chance variables and decision variables. Generally, decision variables are the causes, however, in this framework it is possible as we shall see later to interpret non-decision variables as causes.

By defining a consequence as a deterministic mapping of the states of nature and acts, HS fall in line with Pearl and the counterfactual approach to causality. Through the deterministic mapping, it is possible to evaluate the effect of an action even when such an action did not take place, leading to a counterfactual.

Usually the states of nature are unknown, and thus are the source of uncertainty. They are then defined by the possible combinations of acts and consequences. To clarify this consider the two following examples, in the first, the possible states of nature are known and in the second they are defined by a combination of acts and consequences.

Example 3.4.1 *Sara would like to wear her favourite dress out tonight, but is concerned about the weather as it might rain. The possible states of nature are known, and are rain(R) or no rain (NR). The possible acts are wearing the dress (d) or not wearing the dress(nd). The consequences are wet dress* ***(w)*** *and* ***dry dress*** *(d). The mappings are*

$$\begin{aligned} (R, d) &\rightarrow \mathbf{w} \\ (R, nd) &\rightarrow \mathbf{d} \\ (NR, d) &\rightarrow \mathbf{d} \\ (NR, nd) &\rightarrow \mathbf{d} \end{aligned}$$

Example 3.4.2 *Sara has a headache, should she take an aspirin or not? The states of nature, in this case what led to the headache occurring and whether*

the headache is the type of headache that will go away if an aspirin is taken, as well as other health related variables, are not known. Hence they are defined by a combination of acts and consequences.

State of nature	**act**	
	take	*don't take*
1	gone	not gone
2	gone	gone
3	not gone	not gone
4	not gone	gone

Table 3.3: Example: States of Nature, acts and consequences of taking and not taking aspirin. Gone refers to the headache being gone within 2 hours.

3.4.2 Unresponsiveness

A central concept to HS' decision theoretic approach is that of *(un)responsiveness*, and in particular *limited unresponsiveness.* It is in terms of this concept that HS define cause. The concept of unresponsiveness is similar to the concept of *(ir)relevance* in Pearl (2000) Chapter 7. The idea is expressed in the following way, *if a variable Z is fixed, will altering another, say X change Y?* If the answer is yes, then Y is responsive to X when Z is fixed. If the answer is no, then Y is unresponsive to X when Z is fixed, and Z can be seen as shielding Y from X and thus being a cause.

Basic Notation

The decision problem as seen by HS is described entirely by three sets of variables, the chance variables, the decision variables, the possible states of

the world and a function from the decisions and states of the world to the chance variables.

Let $\mathbf{D}$ be the collection of decision variables, and $\mathbf{U}$ the collection of the chance variables. Individual variables are given by upper-case letters, so for example $Y, X \in \mathbf{U}$ are chance variables, $T, D \in \mathbf{D}$ decision variables. Let $\mathcal{S}$ be the set of possible states of nature. Individual elements of $\mathcal{S}$ are given by s.

Realisations of decision variables are given by lower case variables. For example, t is a realisation of $T \in \mathbf{D}$ and t takes values in $\mathcal{T}$. A realisation of a chance variable $Y \in \mathbf{U}$ when the decision taken is $d \in \mathcal{D}$ and the state of nature is $s \in \mathcal{S}$ is given by $y[ds] \in \mathcal{Y}$ Further, if $\mathbf{V}$ is a subset of $\mathbf{U}$, where $\mathbf{V} = (X, Y)$ then $\mathbf{v}[sd] = (x[sd], y[sd])$ where $x[sd] \in \mathcal{X}, y[sd] \in \mathcal{Y}$ is a particular realisation of $\mathbf{V}$ and we say that $\mathbf{v} \in \mathcal{V}$.

Formally, unresponsiveness is defined as follows. Let $\mathbf{Y}$ be a set of chance variables and $\mathbf{T}$ a set of variables in $\mathbf{U} \cup \mathbf{D}$. We say that $\mathbf{Y}$ is unresponsive to $\mathbf{D}$ in the states limited by $\mathbf{T}$ if for all possible states of the world, if $\mathbf{T}$ assumes the same value for two acts, then so does $\mathbf{Y}$.

Definition 3.4.1 ((Limited) Unresponsiveness) *Given a decision problem described by chance variables* $\mathbf{U}$*, decision variables* $\mathbf{D}$*, and states of the world* $\mathcal{S}$*, and variable sets* $\mathbf{Y} \subseteq \mathbf{U}$ *and* $\mathbf{T} \subseteq \mathbf{U} \cup \mathbf{D}$*,* $\mathbf{Y}$ *is said to be unrespon-*

sive to $\mathbf{D}$ *in states limited by* $\mathbf{T}$*, denoted by* $\mathbf{Y} \not\leftrightarrow_{\mathbf{T}} \mathbf{D}$*, if*

$$\forall s \in \mathcal{S}, d_1, d_2 \in \mathcal{D} : t[sd_1] = t[sd_2] \Rightarrow y[sd_1] = y[sd_2]. \tag{3.17}$$

If $\mathbf{Y}$ *is not unresponsive to* $\mathbf{T}$ *in states limited by* $\mathbf{D}$*, then we say that* $\mathbf{Y}$ *is responsive to* $\mathbf{D}$ *in states limited by* $\mathbf{T}$*.*

For properties of limited unresponsiveness see HS.

If Y is a random variable and D is a decision variable such that $\mathbf{Y} \not\leftrightarrow \mathbf{D}$, then Y is probabilistically independent of D. However, the opposite is not true. This is because in equilibrium situations two variables that are responsive to one another can be probabilistically independent. Limited unresponsiveness and conditional independence are not closely related.

To explain unresponsiveness consider a more complex version of example 3.4.2

Example 3.4.3 *Sometimes Sara has too much to drink and this brings on a very strong headache the day after. When this is the case, taking aspirin never helps. We consider only the states where Sara has a headache to begin with as they show the concept of unresponsiveness sufficiently well. So we say that headache status is unresponsive to aspirin intake in the states limited by Sara having had too much to drink the day before.*

3.4.3 Definition of Cause

HS define *cause* in terms of unresponsiveness as follows:

drink?	**headache?**	*aspirin?*	**gone?**
yes	yes	yes	no
yes	yes	no	no
no	yes	yes	no
no	yes	no	no
no	yes	yes	yes
no	yes	no	yes

Table 3.4: Example: The headache going away is unresponsive to taking an aspirin in the states limited by drink.

Definition 3.4.2 (Causes with respect to decisions) *Given a decision problem described by* $\mathbf{U}$ *and* $\mathbf{D}$ *and a variable* Y *in* $\mathbf{U}$*, the variables* C *in* $\mathbf{D} \cup \mathbf{U} \setminus \{Y\}$ *are said to be causes for* Y *with respect to* $\mathbf{D}$ *if* C *is the minimal set of variables such that* Y *is unresponsive to* $\mathbf{D}$ *in states limited by* C*.*

Going back to example 3.4.3, let the answer to the question *is Sara's headache gone?* be the variable *gone*? $\equiv G$ and let *aspirin*? $\equiv A$ be the variable *take aspirin*, also, let *drink*? $\equiv D$ be the variable representing whether Sara had too much to drink the night before. By looking at table (3.4) we see that whenever the values of D, for the two values of A, the values of G are also the same. For example, if *drink*? $=$ *yes* then for both values of A (*aspirin*? $=$ *yes* and *aspirin* $=$ *no*, G is the same *gone*? $=$ *no* for both) G is a chance variable, and $\{D, A\} \equiv \mathbf{D}$ is the set of decision variables. We can say that G is unresponsive to $\mathbf{D}$ in the states limited by D as when $d[sa_{yes}] = d[sa_{no}]$ then $g[sa_{yes}] = g[sa_{no}]$ for all $s \in \mathcal{S}$, the states of nature. It is also the case that D fulfils the condition of being the minimal set such

that $G \not\leftarrow_D \mathbf{D}$, hence D is a cause of G according to definition 3.4.2.

Chance variables as Causes

By defining cause in terms of unresponsiveness, and allowing variables that limit unresponsiveness to come from the union of the chance and decision variable sets, HS allow for chance variables to be causes. For example, replace *having too much to drink the night before*, D, with *having a migraine headache*, M, in 3.4.3. Then the values of G do not change, and $G \not\leftarrow_M \mathbf{D}$. However M is a chance variable and cannot be intervened upon.

3.4.4 Use of graphical models

The graphical model proposed by HS is not a conventional Bayes net, and is called an *influence diagram in canonical form.* It explicitly represents counterfactual variables by means of *mapping variables.*

In this section we introduce *atomic interventions*, *mapping variables* and *influence diagrams in canonical form.*

An atomic intervention is an intervention variable. It is defined as a variable $\hat{X}$ such that when $\hat{X} = x$ this implies that X is set to x. Refer to them with the intervention variable notation F_X.

The reason that HS introduce the atomic intervention is similar to the reason that the intervention variable is introduced in the decision theoretic framework, namely to be able to be able to encode interventions on observed chance variables. Interventions are decisions, and thus, in HS framework, they are in the set of decision variables $\mathbf{D}$.

3.4.5 Mapping Variables

Each chance variable can be expressed as the deterministic function of a decision variable and the states of the world. Thus, **S** can be interpreted as mapping **U**, the set of decisions to **U(D)**, the set of mapping variables, the possible consequences, realised or not, for the set of decisions **D**. Formally:

Definition 3.4.3 *Mapping variables:*
Given the domain $\mathbf{U} \cup \mathbf{D}$*, and chance variables* X *and* Y *such that for every* $Y \in \mathbf{U} \cup \mathbf{D}$ *there exists* $F_Y \in \mathbf{D}$*, the* ***mapping variable*** $X(Y)$ *is the chance variable that represents all possible mappings from* Y *to* X*.*

To clarify, F_Y is introduced so that it is possible to talk about intervening on Y, and thus Y can be a cause even if it is not a decision variable itself.

With this definition of mapping variables, we see that the chance variables **U** can be then expressed as deterministic functions of the decision variable and the mapping variable. The decision simply selects the appropriate instance of the mapping variable, much in the same way that given a pair of potential responses (Y_1, Y_0), assigning treatment $T = 1$ will reveal Y_1. By defining the mapping variable as above, HS are guaranteeing the existence of a response for every act.

Further, chance variables can also be causes as long as we define what the intervention variable consists of. For example we could say that sex causes breast cancer because we can imagine and express mathematically the idea that if we could change the sex of a person to male at conception then this

would reduce their chance of having breast cancer.

The mapping variable has the additional property of being unresponsive to decisions (see Theorem 3 Heckerman and Shachter (1995)). This again is similar to ignorability assumptions.

Influence diagrams in canonical form

Influence diagrams are DAGs that contain decision nodes. Consider first the influence diagram in figure 3.5. It represents at the relationship between treatment, viral load, health and response in an AIDs/HIV patient. From the

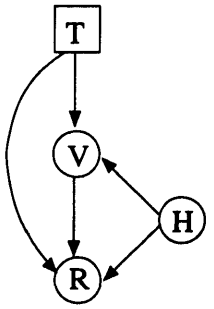

Figure 3.5: The relationship between treatment T, viral load V and response to treatment R and general health H.

influence diagram in figure 3.5 an influence diagram in canonical form can be constructed as shown in figure 3.6. The basic idea is to replace any chance variables with *i)* a mapping variable, which represents the chance element, and *ii)* a deterministic node. The mapping variable takes over the links to other chance variables and has a directed edge into the deterministic node. The edges from the decision nodes into the chance node in the influence diagram go into the deterministic node that replace the chance node. For

details on how to construct such a diagram, refer to Heckerman and Shachter (1995). It is not clear how the mapping variables themselves are related to

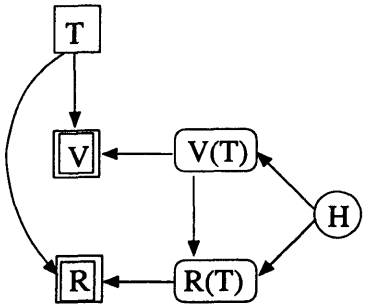

Figure 3.6: The relationship between treatment T, viral load V and response to treatment R and general health H in canonical form with mapping variables taking over the relationships between the variables and the chance variables are deterministic functions.

each other.

Relationship with counterfactual model

This framework is in fact a different but equivalent approach to Pearl's functional counterfactual model. The main difference is that the decisions are the basic blocks in HS. It is not decision-theoretic in the sense of the framework proposed in this thesis.

The correspondence between the counterfactual and HS' framework is easily seen. Given the chance variables $\mathbf{U}$, we can assume that $\mathbf{D}$ contains only the set of intervention variables on $\mathbf{U}$. Let G be the DAG that contains all variables in $\mathbf{U}$ such that $\mathbf{Pa}^{\mathbf{G}}(X) \cup F_X$ are the causes for X with respect to $\mathbf{D}$, where $\mathbf{Pa}^{\mathbf{G}}(X)$ are the parents of X in G. Then the relationships between

$\mathbf{U} \cup \mathbf{D}$ can be expressed by the set of structural equations

$$X = f(\mathbf{Pa^G}(X), F_X, X((Pa^G)(X), F_X))$$

for every $X \in \mathbf{U}$. To clarify consider the following:

1. In Pearl, the observed variables $\mathbf{V}$ correspond to the domain variables $\mathbf{U}$
2. F_X atomic interventions are $do(X)$
3. The graph G is Pearls causal graph
4. Pearls random error U is the causal mapping variable $X((Pa^G)(X), F_X)$.

HS list some advantages of their framework over Pearls; the errors need not be independent as they have a concrete meaning in HS, also by using the canonical form of the influence diagram, it is easier to introduce hidden common causes and their influence on mapping variables. Further it is easier to ascertain the d-separation or independence relationships between different counterfactual variable from the canonical diagram than it is using Pearl's causal DAGs and counterfactual rules.

However, HS's approach is fundamentally deterministic like Pearl and Robin's approaches. The only difference is that the uncertainty here is derived not from random error, but from the unknown underlying state of nature. HS call their framework *decision theoretic*, but it has no decision theoretic elements other than making the decision variable a building block of the framework.

3.5 Discussion

In this section the differences and similarities between the causal inference frameworks covered in this chapter are discussed. We consider first the general assumptions underlying the different frameworks. Then we look at the expressions for causal effects in each framework as a motivation for looking at the assumptions specific to each framework and discuss whether they are justified. Then we look at the similarities between the frameworks. Finally we look at the notation and how it conveys the concepts of interest such as the source of data and whether the problem is predictive or retrospective.

Most of these aspects have been covered in the individual descriptions of the causal models, hence this section aims to clarify and emphasise certain points and will be brief. For a more thorough discussion see the fierce arguments in Dawid (2000).

3.5.1 Assumptions made

There are broadly speaking two types of assumptions made in causal inference. Call them *basic* and *technical.* The former refers to assumptions that determine the structure of the causal inference framework, the latter refers to assumptions made to facilitate evaluation of causal effects. Technical assumptions can be further subdivided into those needed to estimate causal effects from experimental data, and those needed to identify causal effects from observational data.

In this section all three types of assumptions are covered. We will see that

whereas the technical assumptions made in the three types of frameworks discussed so far are often similar, the basic assumptions diverge.

Basic assumptions

The following two assumptions are made in all causal inference frameworks, although the former is made implicitly in the counterfactual and potential response frameworks and explicitly in the decision theoretic framework proposed in this thesis.

Assumption 9 There can be no cause without manipulation

We have seen that Heckerman and Shachter (1995) consider the possibility of variables that cannot be intervened upon as being causes. However, it is necessary to define a potential intervention on them in order to make causal inference.

Assumption 10 Exchangeable treatment units

The units that causal inference will be applied to are exchangeable with the units used to gather the data. Although this is a standard assumption, it is not always clear whether it is appropriate.

Aim of inference and related assumptions

Before looking in more detail at the framework specific assumptions, a quick overview of the causal effects, that is the aim of inference of the different

frameworks.

Aim of inference

Potential responses The individual causal effect for new unit u_{new}, $(ICE(u_{new}))$ of treatment $T = t$ relative to treatment $T = c$.

$$ICE(u_{new}) = Y_t(u_{new}) - Y_c(u_{new}). \tag{3.18}$$

Functional The causal effect of T on Y for the realisation t of T given causal model M

$$p(Y = y|M_t) = p(Y = y|do(T = t), M). \tag{3.19}$$

Where M represents the set of functions that model the variables in cause-effect relationships including an error term.

Generally, the aim of inference is a comparison of a function, such as the expectation of the causal effect of treatment t and that of another treatment, say t^*.

Another point is that although (3.18) is defined as the causal effect, it is often the ICE that is the aim of inference in the functional models framework, as this is just the solution to two different settings of M. Both Pearl and Robins employ counterfactual reasoning and make inference about the ICE.

HS (3.19) is taken as the causal effect by Heckerman and Shachter (1995).

Decision Theoretic The average causal effect of treatment (ACE) $T = t$ relative to treatment $T = c$ for new unit u_{new}

$$ACE = E(Y|F_T = t) - E(Y|F_T = c). \tag{3.20}$$

Where the expectations are taken over the predictive distributions of u_{new} given $F_T = t$ and $F_T = c$ respectively.

Now a look at the specific assumptions made and how they reflect the aim of inference.

Framework specific basic assumptions

Potential responses The values of the potential responses exist for all possible treatments for each unit and is unaffected by actual treatment or how it is administered (see SUTVA assumption). The treatment administered reveals the value of the response to that treatment.

Functional The world around is best expressed in terms of deterministic functions relating effects to their causes and some random disturbance.

HS Each situation is completely determined by the possible states of nature, the possible actions and the possible consequences of the combinations of the states and the actions. In fact, consequences are deterministic functions of the states of nature, and the acts.

Decision Theoretic Causes are decisions.

It is easy to see how the aim of inference drives the assumptions in the potential response framework. If I believe that the ICE is a meaningful quantity and further that it is *the* way of expressing what a causal effect is, then I must make assumptions that turn it into a quantity I can estimate. Hence the existence of potential response assumption in the potential response framework.

If on the other hand I believe that the world is a deterministic machine, then I will define causal effects in terms of functional models as Pearl, Robins and Heckerman and Shachter have done.

Finally, if I believe that the best way to tackle the estimation of causal effects is to treat it as a decision problem and make no additional assumptions about counterfactuals or determinism, then I make the last assumption above.

There are a few additional points worth discussing. Are these assumptions useful? Should the same method be used for predictive, counterfactual and retrospective causal inference alike?

Useful assumptions?

The assumptions made in the counterfactual and potential response frameworks are not useful. In the case of the functional model frameworks the universe is squeezed into a set of deterministic functions, and in the case of the potential response framework, the existence of an infinite number of values that will never be realised is assumed. Neither of these assumptions are

necessary to make inference about predictive problems.

There are three types of questions in colloquial language that involve the concept of cause. These are the predictive *Will my headache go away if I take an aspirin?*, the counterfactual *Would my headache have gone if I had taken aspirin?* and finally the retrospective *My headache is gone, is it because I took an aspirin?*. These three are clearly different types of questions, and should, in my opinion be dealt with using different models.

Retrospective inference

In particular, retrospective questions should be dealt with differently as they have an additional data item in the form of the recorded response of the unit of interest. That is, where the predictive and counterfactual questions are answered using past data only, the retrospective question has the additional $Y_t(u_{new}) = y$ which has already been observed. This requires a different method, unless the treatment unit additivity (TUA) assumption (see item 1 in list 3.5.1) is made. In this case, the new data point does not affect inference about the mean of the distribution of the ACE (given normality assumptions). Dawid (2000) suggests that the reason that this problem is not raised in the literature is because TUA is generally assumed.

Another problem with retrospective inference using or counterfactual or predictive methods is that whilst in the predictive or counterfactual case, we have a clear intervention and a clear effect, in the retrospective case we have a clear effect, but are forcing one past intervention amongst many others to

be the cause.

The decision theoretic framework proposed in this thesis does not go into questions of retrospective causality as they require a new semantic, which might necessarily involve counterfactual variables. Further we pointedly ignore counterfactual questions, considering them the wrong type of question for reasons given in section 2.2.4.

SUTVA and Consistency assumptions

The following assumptions are also made in the potential outcomes framework. They are made to give added structure and avoid complications.

1. Stable unit-treatment value assumption(SUTVA) The values of potential outcomes for each unit are independent of the treatments assigned to other units, and there are no different versions of a treatment.
   This assumption is made to avoid the complications that would ensue if the response to treatment for a unit depended on the treatments other units received, or if different treatments were administered. The SUTVA is violated for example when a member of a household is vaccinated as this has the effect of protecting both him and the other members of the household as there is one less person they can be infected by. It would further be violated if in a clinical trial the group of people who received the drug as opposed to the placebo received two different versions of the drug.

2. Consistency

   The value of the potential response to treatment t for unit u is the same as the realised value if the treatment t is actually administered. $T = t \Rightarrow Y_t = Y$, where Y denotes the realised outcome. This assumption is made to link the counterfactuals to the real world. If we did not believe it the whole potential outcomes/counterfactual framework would fall apart.

The SUTVA has a decision theoretic counterpart called *compatibility* (Dawid 2000): *For two different experimental layouts that both result in the unit u receiving treatment $T = t$, the marginal modes for the response on the unit are identical.* However, it is not necessary, as the framework relates to the real world through the appropriate models. For example, if two treatments are given, then it is possible to code for it by introducing different values for the intervention nodes.

The problem with the ICE

Given a binary treatment and two potential responses for each unit, what can we infer from data about their joint distribution? The answer is not much. We can estimate the marginal distributions of the two responses, but nothing is ever revealed about their joint distribution or their correlation structure. As we are generally interested in making inference about a future unit, the quantity of interest is the individual causal effect (ICE) of u_{new}. We can identify its mean, as this is simply the ACE, however, its variance

will depend on assumptions made about the correlation structure of the two responses. To see this, consider the following simple example taken from Dawid (2000).

Say we have two treatments $T = 1$ and $T = 0$. We assume that the potential responses for these two treatments $Y_1(u)$ and $Y_0(u)$ are jointly iid from a bivariate normal distributions with mean (θ_1, θ_0), same variance σ^2 and correlation ρ. Now, the variance of the ICE is a function of σ and ρ, and cannot be identified from data, unlike the means and the variance, if we do not make further assumptions about ρ. Thus, if two analysts impose different constraints on ρ, the variance of the ICE will be different.

The technical assumptions discussed in the next section are made to overcome this problem.

Technical assumptions

Before discussing the technical assumptions made in the causal models, note that we adopt Pearl's counterfactual notation as opposed to Rubin's. Thus for a binary treatment T where the possible treatments are t, c for treatment and control, the potential responses are $Y_t(u)$ and $Y_c(u)$ respectively for unit u.

Estimation from experimental data

The assumptions covered below refer to situations where experimental data is available. That is, we have randomised treatments under controlled conditions. There follow some simple but very strong assumptions that are not

generally appropriate when dealing with units of people.

Simple Assumptions list 3.5.1.1

1. Temporal stability and causal transience

   (as named in Holland (1986))

   The response to treatment remains the same regardless of when treatment is administered. Also the treatment t does not change the unit u enough to affect the measurement of the response to treatment c later Thus we can identify the ICE by applying both treatments to a unit one after the other.

2. Unit homogeneity

   All the units are identical, thus the responses are also the same for all units.

The above assumptions are generally not appropriate, and some weaker assumptions must be made.

Weaker Assumptions list 3.5.1.2

1. Homogeneity of potential responses

   The potential response pairs $(Y_t(u), Y_c(u))$ are iid given their joint distribution P. This is not the same as the unit homogeneity assumed above, as it is just the responses that must be homogeneous. This assumption is valid if we believe that the units are homogeneous enough and the treatments act on these in a very similar way. For example, this may be a plausible assumption if we are administering a drug to a

group of rats which are all related and identically bred and raised up to the point of the trial. If we have two groups of rats, one lab bred and the other sewer bred, we may not be happy to make this assumption.

2. Treatment unit additivity (TUA)

   The ICE is the same for all units. Denote the ICE by τ, then $Y_t(u) - Y_c(u) = \tau$ for all u. So we can infer the non-realised response simply by subtracting τ from the realised response. This assumptions is a weaker version of the unit homogeneity assumption and is implied by it. It corresponds to $\rho = 1$ in the above discussion of the ICE. This may be an appropriate assumption for the lab/sewer rat trial if we believe that the rats on average would react in the same way, except that the sewer rats are slightly less (or more) healthy than the lab rats, and thus the ICE would remain the same.

3. Monotonicity

   The response to treatment is always larger (or smaller) than the response to the control. This assumption makes sense for binary treatments or multiple treatments that have a natural ordering such as escalating dosages. Formally: $Y_t(u) \geq Y_c(u)$.

None of these assumptions need to be made in the decision theoretic framework as the the target of inference is the ACE, which depends only on the marginal distributions of the observed responses.

Identification from observational data

The biggest problem in the area of identification of causal quantities from observational data is that the assumptions that allow this are not often made explicitly. There follow two which are.

1. Rubin - Invariance to treatment assignment

   It says that the way the unit responds to treatment, i.e. the value it assumes does not depend on how the treatment was administered. This assumption basically says that we can make causal inference from observational data without further ado.

2. Robins - No unmeasured confounders The counterfactual Y_t does not depend on the current treatment given past treatments and covariate information. It is equivalent to saying that a treatment is assigned at random given the past.

   $$Y_t \perp\!\!\!\perp T | \text{past treatment and covariate history}$$

   This assumption is analogous to Rubin's *conditional ignorability* assumption, which holds if there is a set of variables Z say, such that $Y_t \perp\!\!\!\perp T | Z$.

The decision theoretic model again makes none of these assumptions. Without them it is still possible to make inference on matters of interest. For example Dawid (2002) covered the problem of partial compliance without the use of counterfactuals. The problem of direct and indirect effects as well as the effect of treatment on the treated are tackled in non-counterfactual terms in this thesis in chapters 5 and 6 respectively.

3.5.2 Translations

How the counterfactual frameworks translate into the decision theoretic framework will be explored at length in the remainder of this thesis, here we focus briefly on how the potential outcomes framework translates into the functional model framework.

The basic difference between the potential outcomes framework in section 3.2 and the functional model framework in 3.3 is that the former takes the potential outcomes $Y_t(u)$ as primitives and defines the rest accordingly (and thus must make assumptions such as the consistency assumption in list 3.5.1), whereas the latter takes the functional model and the $do(\cdot)$ as the primitives and defines the potential responses as solutions to the structural equations given an action given by a setting of the $do(\cdot)$ operator.

Thus if we take the potential response $Y(t,u) = Y_t(u)$ this is equivalent to $Y_{M_t}(u)$, the unique solution to the set of equations given by the model M under the intervention $do(T = t)$, denoted by M_t.

Another interesting difference between the two is how the concept of randomness is introduced. In the potential outcomes framework, the random element comes from the randomness of the units. So, the value of $Y_t(u)$, given by Y_t is a random variable. In the functional model framework, the randomness comes from the set of unknown background or error variables $\mathbf{U}$. These are essentially the same as the units in that these background variables represent all the unknown components any unit may have. In the case of units being people, U is everything up to the unique genetic make-up of

each individual.

There is an interesting consequence of how the two frameworks are defined that shows that although most aspects can be translated from one to the other, one cannot.

In Rubin's framework, the assumption of no confounders, translated into the decision theoretic notation, is given by

$$Y_1 \perp\!\!\!\perp F_T \text{ and } Y_0 \perp\!\!\!\perp F_T. \tag{3.21}$$

In Pearl and Robins' frameworks it is given by

$$(Y_1, Y_0) \perp\!\!\!\perp F_T. \tag{3.22}$$

The former states that the potential responses are each marginally independent of the intervention variable, and the latter states that the potential responses are jointly independent of the intervention variable. These are not the same from a probabilistic point of view although the latter implies the former. Further, if we believe (3.21) and also that $(Y_1, Y_0) \not\perp\!\!\!\perp F_T$, that is that the potential responses are each marginally independent of F_T but jointly dependent of F_T, then this cannot be described graphically. Clearly this is a strange circumstance to imagine[2] , and it is usual to assume both (3.21) and (3.22).

In section 3.4.5 we see how HS' framework is an extension of the functional model. Although HS take the concepts of states of nature, actions and consequences as primitives, the framework developed is essentially a deterministic

[2]Although I have it from personal communication with A.P.Dawid, that this makes sense in dynamic situations.

functional model. In it, the source of uncertainty comes from the unknown states of nature, not from random error or unknown background variables as it does in Pearl's version, or units as in Rubin's version.

Which framework is a more appropriate description of causality is argued in sections 3.6.3 and 7.4.4 in Pearl (2000), Rubin (2004) and in Heckerman and Shachter (1995) and is beyond the scope of the current discussion.

3.5.3 Notation

It has struck me over the course of my research that notation is an essential element of our understanding of a problem. The same expression can be obscure in one notation and clear in another. Whether an expression is clear can depend on training; however, some notations are more prone to being misinterpreted than others.

Take the following two expressions, one is in counterfactual terms and the other is in decision theoretic terms: *(i)* $p(Y_t = y) = a$ and *(ii)* $p(Y = y|F_T = t) = a$! The latter says that the probability of $Y = y$ given intervention has taken place at T, setting it to t, is a. The former says that given $T = t$, the probability of $Y = y$ is a, that is $p(Y = y|T = t, F_T = \emptyset) = a$. If we further assume that $Y \perp\!\!\!\perp T|F_T$, then *(i)* says the same as *(ii)* without difference in the expression itself.

Further, *(i)* can also refer to a hypothetical variable whose realisation we shall never be able to see, something that cannot be expressed in the decision theoretic framework. As the decision theoretic framework codes all

interventions explicitly, it is harder to get confused by the notation and led astray.

3.6 Look ahead and Conclusion

We have covered the decision theoretic approach in Chapter 2 and looked at the competing causal models in this chapter. The remainder of the thesis looks at aspect of causal inference in the decision theoretic framework. In Chapter 4 we look at the assumptions underlying causal inference from observational data when using *causal discovery algorithms*. These assumptions turn out to be very strong and rarely justified. In Chapter 5 we tackle the problem of direct and indirect effects in non-counterfactual terms, and see that in the decision theoretic framework these effects are simpler to express and manipulate by introducing fictional variables. Finally, in Chapter 6 we look at how to estimate causal effects when randomised trials are not conducted and the only data available to us is data on the treated. The quantity we try to estimate is the effect of treatment on the treated.

Finally a quote from Lauritzen (2004).

> ...I see the different formalisms as different languages...and I have no difficulty accepting that potential responses, structural equations, and graphical models coexist as languages expressing causal concepts each with their virtues and vices.

Lauritzen's open minded view is in stark contrast with the attitudes in the

causal inference literature. Although I believe that the decision theoretic approach is the most appropriate for predictive questions, it may be the case that questions of attribution need a new semantic that includes the use of counterfactuals. It will be up to the reader to conclude which of the frameworks he or she favours.

Chapter 4

Causal Discovery Algorithms

4.1 Introduction

A contended issue in causal inference is that of *causal discovery.* Large databases and increasingly efficient computers have prompted the development of algorithms that aim to extract causal relationships from observational data. Such causal discovery algorithms have been put forward by Pearl and Verma (1991), Cooper (1997), Silverstein et al. (2000), Heckerman et al. (1999) and Spirtes et al. (2000) (see also the web based Tetrad Project at www.phil.cmu.edu/projects/tetrad) amongst others.

Very strong assumptions underpin the process of causal discovery from observational data. The aim of this chapter is to clarify these assumptions. In particular, the assumption referred to as the *fundamental assumption* is the focus of this discussion, as it allows the transition from association to causation, and is not often addressed explicitly in the causal discovery literature.

A slightly different but equally sceptical view of the validity of causal discovery algorithms is given in Freedman and Humphreys (1999). The paper also focuses on the problem of inferring causation from association. It argues that it is not reasonable to expect an automated process to be able to distinguish between a causal relationship and an association, as it is complicated process that takes a lot of thought on the part of human researchers. Further the software developed by Spirtes et al. (2000) is tested and revealed to be flawed. The discussion in this chapter makes a formal mathematical distinction between causation and association by introducing the intervention node, thus making the difference more immediately visible.

Section 4.2 motivates the discussion of causal discovery from observational data by giving a simple example of the process. Section 4.3 describes the basic structure of a constraint-based algorithm. Sections 4.4 and 4.5 state the fundamental assumption in terms of simple causal DAGs and then in terms of augmented DAGs. Section 4.6 details other assumptions underlying the causal discovery process. Section 4.7 gives an example of what inference could be drawn if the assumptions are deemed to hold. Concluding remarks are made in section 4.8.

4.2 Motivating Example

Consider the following example based on Silverstein et al. (2000). It is a simplified case of the *market basket problem.* A *basket* is a boolean vector assigned to each customer with an entry for every item in the market. The

entries can be either 0 or 1 depending on whether the customer bought the item or not respectively.

For example, a basket representing the contents of a particular customer's shopping trolley in a supermarket is a vector that indicates the absence or presence of every item in the supermarket. If the supermarket sells only cereal, burgers, ketchup and milk, the contents of one particular shopping trolley are represented by the vector $b = (1_C, 0_B, 0_K, 1_M)$, meaning that the customer bought cereal and milk but not burgers or ketchup.

Consider a company that sells CDs online. It assigns a basket to each new customer. The sales director is interested in finding patterns in the shopping tendencies of customers. To do this, he runs market basket data through a causal discovery algorithm. As he interprets the results as causal, these tell him that buying CDs by established morbid metal band *Blame the living* **causes** customers to buy CDs by new morbid metal band *Kings of the dead.*

As a consequence, the sales director decides to double the price of *Kings of the dead* CDs and simultaneously discount the *Blame the living* CDs by 25 %, thinking that company will make a profit. Unfortunately, the *Blame the living* CDs sell out and the *Kings of the dead* CDs stay on the shelves.

The picture so far is the following. *Observational* data on a set of variables was analysed and the trends found were interpreted as causal. These were then used to *intervene* on the observed variables to obtain a desired result which however, did not materialise. In other words, a system in its natural state was assumed to behave in the same way as it would under intervention.

This does not necessarily have to be the case as seen in the above example.

4.3 Causal Discovery Algorithms

There are principally two approaches to causal discovery, one is that of *constraint-based* discovery algorithms and the other is that of *Bayesian* discovery algorithms. This section describes the simplest type of constraint-based algorithm and goes briefly into how it differs from the Bayesian approach.

4.3.1 Simple constraint-based algorithm

Consider a set of observed variables $\mathbf{V}$ for which there is a large amount of observational data. The causal discovery algorithms aim to find the causal relationships among the observed variables $\mathbf{V}$.

The steps of the simple constraint-based algorithm are the following; first the data is tested to find a set of conditional independence relationships using standard statistical tests such as the χ^2 test. Next, these conditional independence constraints are used to generate a set of Markov Equivalent DAGs, that is, DAGs that are indistinguishable with respect to the set of conditional independence relationships (see definition 2.4.4). If there is prior knowledge about some relationships, such as precedence, these can be used to rule out some of the Markov Equivalent DAGs.

If the set of Markov Equivalent DAGs (minus those ruled out by external constraints) have directed edges in common, these are interpreted as causal.

It is worth emphasizing again, that although the common directed edges are used for causal inference, the procedure that generated the graphs used only observational data.

The more complex algorithms take into account the possibility of unobserved common causes leading to what may appear to be causal links between two or more variables. These are not discussed in this paper for the sake of simplicity and as they do not change the main argument.

For a full description of different types of discovery algorithms see Silverstein et al. (2000), Pearl and Verma (1991), Spirtes et al. (2000) Chapter 5, Cooper (1997) , Pearl (2000) Chapter 2 for constraint-based algorithms and Heckerman et al. (1999) for Bayesian algorithms amongst others.

Constraint-based vs Bayesian

The most important difference between the constraint-based approach and the Bayesian approach for the purposes of the current discussion is that whereas the constraint-based method assumes the conditional independences found from the data using the χ^2 tests are "true" and uses these directly to determine the set of Markov Equivalent DAGs, the Bayesian approach attaches uncertainty to them. See appendix B for a more detailed explanation of both constraint based and Bayesian causal discovery algorithms.

4.4 Fundamental Assumption

The fundamental assumption needs to be made before the algorithms are initiated or even programmed. Making this assumption allows the transition from statistical inference using observational data to inference about interventions.

It can be expressed as follows:

We assume that there exists a unique causal DAG that describes all possible regimes involving the observed variables (this may include their connection to possible unobserved common causes as detailed in section 4.6). This means that any experiment we chose to consider, involving the observed variables, where some or all of these are intervened upon, is described by the same causal DAG. Further, the observational case is also described by the same causal DAG.

We are therefore assuming two things. The first is that there is an underlying causal structure between the observed variables. That is, the relationships between the variables are *a)* not the product of an exceptional and rare combination of information, and *b)* not the result of a dependence that is not causal in nature. The second is that the causal and the natural structures can be summed up in a single DAG. Both are very strong assumptions. They can be tested only if every possible experiment on the set of observed variables is carried out. This may be either impossible or unethical even when dealing with a small set of variables.

Given that we accept the fundamental assumption, and the causal discovery algorithm works, it follows that exactly one of the Markov equivalent DAGs found by the algorithm does not just graphically code the conditional independence relationships between the observed variables, but *is* the unique causal DAG and can therefore be given a causal interpretation.

4.4.1 Examples

The following examples are based on constraint-based algorithms as these are simpler to understand and do not differ from Bayesian algorithms with respect to the fundamental assumption.

Say we have some observational data over a finite set of variables $\mathbf{V}$. Let the set of conditional independence relationships between the elements of $\mathbf{V}$ (or alternatively the joint distribution of the elements of $\mathbf{V}$) in the observational situation, be known. The following simple example explains how the casual discovery works and what it assumes. Note that we do not take into consideration the possibility that the results of the χ^2 tests are wrong and that we have extracted a false conditional independence from the data.

Example 4.4.1 *Let* $\mathbf{V} = \{X, Y, Z\}$ *be a set of observed variables. Observational data involving these three variables is analysed using* χ^2 *tests and the following conditional independence relationship is found,*

$$X \perp\!\!\!\perp Z \,|\, Y.$$

This is consistent with the three DAGs in Figure 4.1.

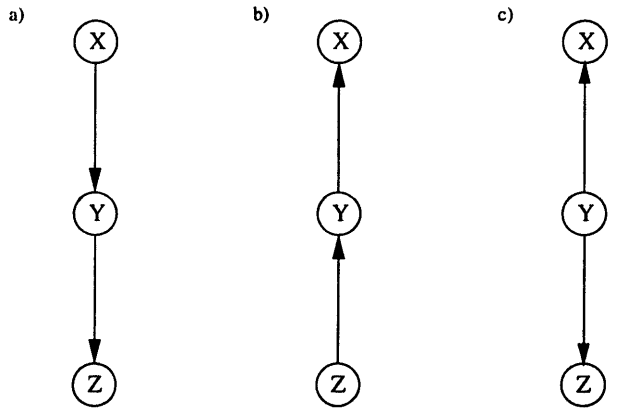

Figure 4.1: These three DAGs are Markov equivalent, each encoding the conditional independence $X \perp\!\!\!\perp Z \mid Y$

The *causal* interpretation of DAG *a)* states, for example, that intervening to change the value of Z would not affect the probability distributions of X or Y, nor the conditional distribution of Y given X; further, the distribution of Z, given the value of Y, would be unaffected by an intervention to change the value of X.

In general, let $\mathcal{D}_V$ be the set of Markov equivalent DAGs over the vertex set $\mathbf{V}$ found using the discovery algorithms on the observational data. By the fundamental assumption, we assume that there exists a unique causal DAG, D_V, on the same vertex set $\mathbf{V}$, that underlies all experimental and observational situations. It follows again from the fundamental assumption, that there exists a graph $G \in \mathcal{D}_V$ such that G is identical to D_V. That is, one of the DAGs in $\mathcal{D}_V$ represents the causal structure between the observed

variables.

In example 4.4.1, the assumption allows us to state that one of the three DAGs in Figure 4.1 represents the underlying causal structure of the three variables. As it stands, this only tells us that there is no causal relation between X and Z that is not mediated by Y. In particular it is not possible to estimate any causal effects without making further assumptions or using prior knowledge, such as temporal ordering, to impose additional constraints.

To clarify the above, consider the example given in section 4.2 in more detail.

Example 4.4.2 *The CD company collects data on sales and prices on a weekly basis over the course of a year. The prices are standardised according to economic factors such as inflation, so although the price in US dollars of a CD remains \$14.99, the standardised price varies naturally over the course of the year*[1].

The customer baskets and the sales data are run through a causal discovery algorithm and result in the DAG in Figure 4.2. C_K *is the cost of the CD by*

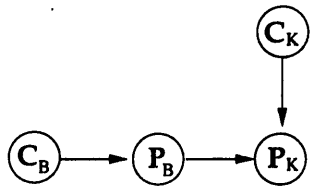

Figure 4.2: Discovered DAG relating cost and two music CDs.

new morbid metal band Kings of the Dead. P_K *is the number of customers*

[1]The standardisation is necessary in order to introduce variation into the price of the CDs

visiting the website who buy the CD. C_B is the cost of the CD by established morbid metal band Blame the living, *and P_B is the number of customers visiting the website that buy the CD.*

For the sake of simplicity we do not take into account economic forces such as demand affecting supply and cost etc. This means that we can exclude the only other DAG that is Markov Equivalent to the DAG in Figure 4.2, which has an arrow pointing from P_B to C_B.

The sales director of the company interprets this DAG as causal and intervenes on the price of the CDs. This results in a loss as explained in section 4.2. This is because the DAG does not encode causal relationships but associations, which are not necessarily invariant under intervention.

An explanation for why this graph was generated (instead of another) might be that there were unobserved common causes. For example, both the cost of the CDs and the number of people buying it may have been influenced by temporary fashion trend. Alternatively, the graph was generated because there was a non-causal dependence between the variables or simply by chance.

4.5 Fundamental Assumption in terms of the Augmented DAG notation

The assumption can be expressed in the following way by using the augmented DAG notation.

There exists a unique augmented DAG A_V that describes all possible situations arising from experiment as well as the observational case on the set of

observed variables. Hence one of the Markov equivalent DAGs discovered by the algorithm is the core DAG of A_V and therefore codes causal relationships.

The augmented DAG notation clarifies the assumption, as it explicitly codes interventions as nodes in the graph. The DAGs the algorithms discover represent the core of augmented DAGs with the intervention nodes taking on $\emptyset$ values. Under the fundamental assumption, exactly one of them also represents situations in which intervention takes place.

4.5.1 Examples

Consider as before the finite vertex set $\mathbf{V}$. Further, let A_V be the unique augmented DAG that underlies all possible experiments as well as the observational case and let $F_V = \{F_X : X \in \mathbf{V}\}$ be the set of intervention nodes in A_V. Also let D_V be the core DAG of A_V. The assumption is now that there exists a DAG $G \in \mathcal{D}_V$, the set of Markov equivalent DAGs of D_V such that G is identical to D_V.

Example 4.5.1 *Let the set up be the same as in Example 4.4.1, that is, let $\mathbf{V} = \{X, Y, Z\}$ be a set of observed variables. As in the previous case, analysis of observational data has resulted in the following conditional independence*

$$X \perp\!\!\!\perp Z \,|\, Y.$$

The augmented DAGs corresponding to the above circumstance are given in Figure 4.3.

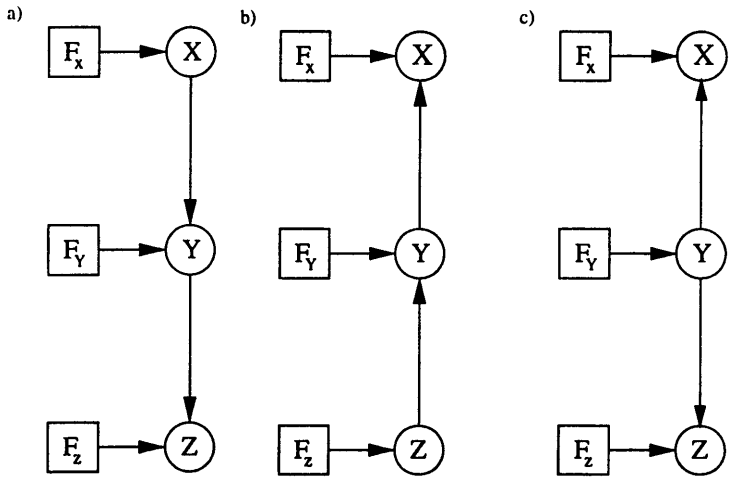

Figure 4.3: These graphs represent the interventions at each node

It is important to point out that although the augmented DAGs in figure 4.3 have intervention nodes at every observational node, this need not be possible in practice. Some nodes such as those representing sex or age cannot generally be intervened upon in real situations.

As the conditional independence DAGs found by causal discovery algorithms are interpreted as causal, they can be extended to augmented DAGs by adding intervention nodes. These augmented DAGs can then be used to describe the causal relationships between observed variables.

Compare for example DAGs *a)* in figures 4.1 and 4.3. If we accept that figure 4.1,*a)* contains causal information, and that therefore figure 4.3 *a)* represents the causal relationships explicitly via the intervention nodes we can read the following off figure 4.3 *a)* using the moralisation criterion (2.4.3):

(i) a change in the value of Y will not affect the distribution of X.

(ii) Z is not independent of intervention on Y (given that the graph is faithful, see assumption 4.6.3) as it is a descendant of F_Y. Finally,

(iii) *given* the value of Y, Z is independent of whether Y arose naturally or by intervention. That is, $Z \perp\!\!\!\perp F_Y | Y$.

Although item *(i)* can be read off DAG *a)* in figure 4.1, items *(ii)* and *(iii)* can only be read off DAG *a)* in figure 4.3 as they refer to the relationships interventions have with the variables in the problem. Thus they are valid only if the fundamental assumption is made.

Example 4.5.2 *Recall example 4.4.2 in the previous section. The DAG in Figure 4.2 is the core of the augmented DAG in Figure 4.4 omitting intervention nodes on the number of customers buying the CDs as we assume that it is not possible to force a customer to buy a product. The DAG can now be inter-*

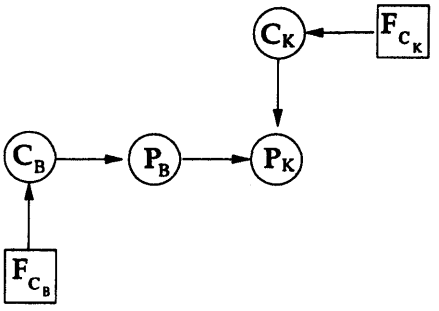

Figure 4.4: Figure 4.2 as an augmented DAG.

preted as follows. Intervening to change the cost of the CDs (within certain limits), will not change the relationships between the 4 observed variables,

as the DAG represents causal relationships which are invariant to external manipulation.

This clarifies the concept of intervention as an additional aspect of the problem that is not inherently obvious in the causal DAG in the previous version of the example.

A further point is that the augmented DAG 4.4 is no longer Markov equivalent to any other DAG as we exclude the possibility of reversing the arrows from the intervention nodes to the chance nodes.

4.6 Further Assumptions and Conditions

The following section lists some of the principal assumptions made either explicitly or implicitly in the literature, in particular for constraint-based algorithms. Not every assumption is clearly specified in all papers, although most are mentioned. These assumptions are important in their own right. However, they do not have the same importance as the fundamental assumption as they are either conventions, or of a more technical nature.

4.6.1 Markov Condition

A node must be independent of its non-descendants given its parents.
This condition is a semantic requirement of graphical modelling, without it, using graphs to represent causal structure makes no sense.

4.6.2 How can 2 associated variables be causally related

This assumption states that there are only three ways in which observed variables that are found to be associated in the data can be related: either one is the cause of the other, or both have at least one common cause in the form of a third unobserved variable (Pearl 2000 Chapter 2). The relationship is shown in Figure 4.5. The arrows in the DAGs in Figure 4.5 represent causal relationships. This assumption is made to limit the number of ways in which

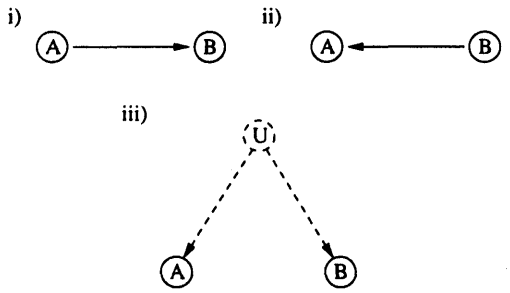

Figure 4.5: The three ways two observed variables can be causally related if they are statistically dependent.

the observational data can be related to the underlying causal structure. It is an important assumption and is very closely related to the fundamental assumption. However, it is not assumed in simple constraint based algorithms or Bayesian algorithms that do not consider the possibility of latent variables. This makes such simple algorithms even more restricted, as this means that whenever two variables are related it is assumed that one causes the other.

It is a necessary assumption when the algorithms include latent variables as these must be related to the observed variables in some way. However, it

need not be true. If we take the stance that causation can only be inferred from intervention, there is no reason to believe that variables found to be associated in observational data are causally related, either directly or through a common parent. The association may be spurious or have a non-causal source.

It is possible to test this assumption if we run experiments to determine whether the postulated causal relationships exist. However in the current context where we are using observational data, precisely because there is no available experimental data, the assumption cannot be tested.

4.6.3 Faithfulness

An independence relationship is implied by the Markov Condition applied to a DAG if and only if it also holds in the associated probability distribution. This condition is imposed to exclude the possibility of negative and positive correlations cancelling. (Spirtes et al. 2000).

4.6.4 No selection bias

The data is a sample drawn at random from the population.

4.6.5 Asymptotic properties of statistical tests

The statistical tests performed to find the conditional independence relationships must be valid, that is, the limiting behaviour of the tests, as the sample size tends to infinity is to find the "true" conditional independence relationships.

Assumptions 4.6.3 and 4.6.4 are made to exclude exceptional data sets, although it is worth noting that while assumptions 4.6.4 is in principle testable, assumption 4.6.5 is not. Also assumptions 4.6.4 and 4.6.5 have no specific relation to the study of causality.

4.6.6 Database completeness and discreteness of Variables

Some of the simpler algorithms, such as Cooper (1997), require the databases to have no missing values and for the variables to be discrete.

4.6.7 Existence of Root variable

Some algorithms Cooper (1997), require that there exist a known variable W, chosen according to background knowledge in the set of observed variables $\mathbf{V}$, that has no causal parents in $\mathbf{V}$. This variable must exist and must be specified before the algorithm is run. Typical examples of such a root variable are age or race as these are generally variables that cannot be intervened upon.

Conditions 4.6.6 and 4.6.7 are not necessary and are made to facilitate the algorithms. They can be excluded in more complex algorithms.

Finally, a technical point. From the point of view of the algorithm all nodes can at least potentially be intervened upon as the algorithm makes no distinction between pure chance nodes and nodes that could be decision nodes. However, there are nodes that can in practice not be intervened upon in the context of the experiment. For example, in Examples 4.4.2 and 4.5.2, the variables P_B and P_K could not be intervened upon as it is impossible

(bar advertisement which I did not take into account for this simple example, or bodily damage) to force a customer to buy CDs. It should be up to the analyst to determine when it is sensible to consider a node as a potential decision node. This will also depend on how they chose to interpret causality. Some will consider gender a node that can be intervened upon at least in a hypothetical sense (Heckerman and Shachter 1995), others, in particular the author, will exclude such a possibility entirely.

4.7 What can be inferred from the discovered DAGs?

Given that we are prepared to accept the fundamental assumption and consider the discovered DAGs as representing causal relationships, what type of inference can we make?

In general, if we have a set of Markov equivalent DAGs $\mathcal{D}_V$ over a set of variables $\mathbf{V}$, these can sometimes also be represented by a partially directed graph D_* in the following way: if X , $Y \in \mathbf{W}$ then there is an arrow from X to Y in D_* if and only if there is an arrow from X to Y in every DAG in $\mathcal{D}_V$. The links between nodes which have different directions in different elements of $\mathcal{D}_V$ are left as undirected edges in D_*. The arrows indicate direct (with respect to $\mathbf{V}$) causal relationships between the variables they connect.

Example 4.7.1 *In example 4.4.1, assume that Y precedes Z in time. Then figure 4.1 b) can be excluded as a possible causal interpretation of the data. In augmented DAG terms, we are left with the two graphs on the left of figure*

4.6. These can be reinterpreted as the partially directed graph on the right side of figure 4.6, where the dashed line means that the direction of that arrow is not known.

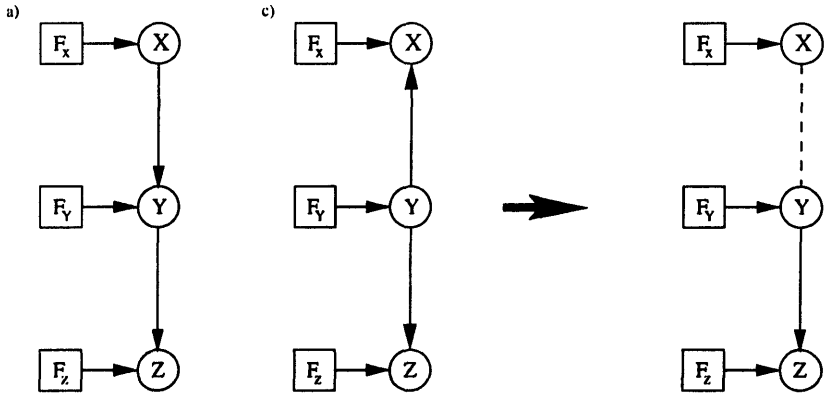

Figure 4.6: In this case, we can say that changing Z will not affect X or Y, that whether Y is set or arises naturally will not affect the distribution of Z given Y, but nothing can be said about the relationship between X and Y

Looking at the augmented DAGs in Figure 4.6, there are two types of inference we can make. The first type is structural. For example, we can see that changing the value of Z will not affect the distributions of Y or X, as $X \perp\!\!\!\perp Z|Y$ and further Y is a parent of Z. We can also read off that how the value of Y arises (via intervention $F_Y = y$ or naturally $F_Y = \emptyset$) does not affect the conditional distribution of Z given Y, as $Z \perp\!\!\!\perp F_Y|Y$. Note that we can read such relationships from the graph because we attach intervention nodes in a very specific way, that is, each chance node is assigned its own intervention node, which is not linked to any other chance node.

The second type of inference is the estimation of the causal effect of an intervention. The causal effect of setting $Y = y$ ($F_Y = y$) on Z can be easily estimated from the data in the following way: First consider what is meant by the *causal effect* of Y on Z:

$$\begin{aligned} p(Z|F_Y = y) &= \sum_Y p(Z \mid F_Y = y, Y)p(Y \mid F_Y = y) \\ &= p(Z \mid F_Y = y, Y = y) \end{aligned} \tag{4.1}$$

that is, the distribution of Z given Y is set by intervention to some value y. Note that $p(Y \mid F_Y = y) = 1$ when $Y = y$ and 0 otherwise by the definition of F_Y. Now, from Figure 4.6 we can see that

$$Z \perp\!\!\!\perp F_Y \mid Y;$$

hence Equation 4.1 can be written as

$$p(Z \mid F_y = y, Y = y) = p(Z \mid Y = y). \tag{4.2}$$

This conditional distribution is estimated from the observational data given that $Y = y$ has been observed.

4.8 Conclusions

If the fundamental assumption and additional assumptions are considered to hold, the discovery algorithms can be used to determine causal relationships from large databases in medicine, economics and the social sciences where

these are not evident. They have already been applied to medical and social science data (Spirtes et al. 2000 and Cooper 1997) with varying degrees of success.

However, the fundamental assumption itself, cannot be verified unless experiments are run, and therefore, inference based on causal discovery algorithms must be used very carefully.

The discovery algorithms could be used as exploratory studies in cases in which manipulation of a particular variable is difficult or costly. Although a relationship between two variables in a discovered DAG would not guarantee a relationship under experimental conditions, it might indicate the possibility of such a relationship.

It would be of interest to explore whether these causal discovery algorithms would find causal relationships in complex experimental situations. For example in the case of black box interventions, where a system with many variables in equilibrium is disrupted by an intervention, such an algorithm could perhaps be useful. An example would be policy interventions such as introducing the congestion charge in central London.

Another avenue of research that has recently emerged is that of *multiple-bias models* (Greenland 2005), which models the discrepancy between experiments and observational studies as bias parameters using fully Bayesian methods or empirical approximations to it.

Chapter 5

Direct and Indirect Effects

5.1 Introduction

Direct and indirect effects are a common concept in the social sciences, where SEMs are used to illustrate and evaluate causal relationships. The problem has also been tackled in the causal inference literature, starting with Robins and Greenland (1992), and more recently, Pearl (2001b), Robins (2003) and Rubin (2004). Although the problem is tackled using counterfactual or potential response methods by all of the above, their approaches vary, as different initial assumptions are made.

It is the aim of this chapter to look at the problem in non-counterfactual terms. This approach requires fewer assumptions than the counterfactual counterparts in order to identify similar quantities and is more versatile. We take the paper by Pearl *Direct and Indirect Effects* (2001) as a starting point.

Note that there are many different concepts referred to as direct and indirect effects. This chapter explores a specific subset of problems that fall

under this heading which can be described informally as follows. A treatment is administered and a response is recorded. However, there exists a variable that is thought to mediate the effect of the treatment on the response, in some way *channelling* a part of the treatment effect on the response. Sometimes we are interested in the effect of the treatment that is not mediated, the *direct* effect, and at other times we are interested in the mediated *indirect* effect.

Section 5.2 illustrates the problem and why it is of interest using examples from different sources. Section 5.3 looks in detail at the paper by Pearl (2001b), how direct and indirect effects are defined and what criteria are necessary for identification of these quantities under experimental conditions. Section 5.4 describes in detail the decision theoretic framework for the expression of direct and indirect effects. Manipulation variables that code randomised interventions, are introduced. Section 5.5 tackles the problems of identifying said effects. Section 5.6 extends the framework by adding a new variable that changes the definition of the manipulation variable. Finally, section 5.7 suggests some further extensions.

5.2 Examples of Direct and Indirect Effects

There follow examples of direct and indirect effect problems drawn from diverse sources. Some are taken from Pearl (2001b), others are taken from marketing and medical literature. The final example considers the problem of surrogate markers, a subject dealt with in non-graphical terms in Rubin (2004).

The examples that follow will allow the reader to form an idea of what we mean by direct and indirect effects before we proceed to formalise the framework. Each example will be accompanied by a DAG to provide a visual point of reference.

Example 5.2.1 *Treatment with headache side-effect*

A drug treatment has headaches as a side effect. Patients who suffer from these headaches tend to take aspirin to alleviate it, and it is thought that the aspirin may have an effect on the response to the drug treatment. The drug company is interested in both the direct effect of the treatment on the disease as well as the indirect effect the aspirin may be having on the treatment response. Both effects are considered stable physiological effects. A graphical representation is given in figure 5.1.

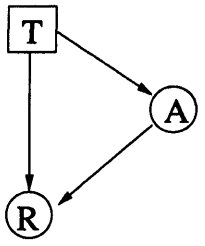

Figure 5.1: Graph representing relationships between treatment T, response R and aspirin A. T is in a box as it is a decision variable.

Example 5.2.2 *Birth-control Pill*

A birth-control pill is suspected of causing thrombosis. However, as pregnancy also gives rise to thrombosis, and the birth-control pill reduces the likelihood

of pregnancy, it has an additional negative indirect effect on the occurrence of thrombosis.

The pharmaceutical company that produces the birth-control pill is interested in the direct physiological effect of the pill on the occurrence of thrombosis. This is considered a stable relationship which can be measured. The indirect effect on the other hand, involves pregnancy and will therefore depend on socio-economic factors such as religion and marital status amongst others, which cannot be controlled for.

The DAG in figure 5.2 is a graphical description of the problem.

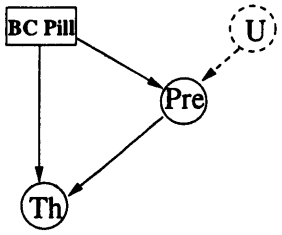

Figure 5.2: The relationship between treatment *BC Pill*, response *Th* and intermediate confounder *Pre*, which may itself be influenced by unobserved factors *U*.

The next example comes from MacDonald and Smith (2004) in the management literature.

Example 5.2.3 *Technology Mediated Communication*

A suppliers' association commission is interested in whether the introduction of technology mediated communication (TMC) such as video conferencing and interactive websites is benefitting their buyers' future intentions (FI).

They believe that introducing TMC has a positive direct effect on FI, as well as an additional indirect effect via trust (T), commitment (C) or both.

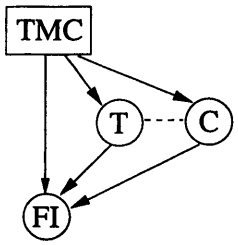

Figure 5.3: The relationships between TMC, T and C, and FI. The edge between T and C is dashed and undirected because the nature of the relationship between these two variables is not known.

This example is slightly different as it has two mediating variables. These can be taken as one and the framework applied.

The following example comes from Solanki et al. (2000) in the health services literature.

Example 5.2.4 *Effect of cost-sharing on utilisation of preventive medical services*

Many Health insurance companies implement cost-sharing schemes. These are provisions of health insurance companies that require the insured to cover some of the costs incurred by the medical service. It is feared that these schemes have a negative effect on the reception of preventive services. The use of preventive medical services occurs in two ways. Either the patient actively seeks out the preventive service (such as a Pap smear or a mammogram) or they are referred to a preventive service as a consequence of a visit to the

GP for a general check-up. Hence cost-sharing has a direct effect on receiving preventive services by discouraging patients to actively seek out the service and an indirect effect by discouraging patients to go for check-ups, as neither service is free. See figure 5.4

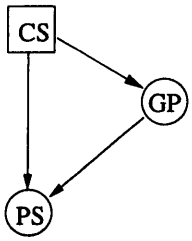

Figure 5.4: Cost sharing schemes have a direct effect on the number of people receiving preventive services (PS) and and indirect effect by reducing the number of visits to the GP (GP).

It is important to be able to identify the two effects and determine their magnitude so as to be able to implement policy changes. For example, by encouraging patients through incentives to visit their GP it would be possible to eliminate or at least control part of the indirect effect.

The final example expresses the concept of surrogate markers in terms of direct and indirect effects. The basic setup is as follows. If T is a treatment variable, and R a response that is either difficult to observe or needs to be avoided, then a surrogate marker S is a variable that can predict R well for any interesting value of T and that "shields" R from T as much as possible.

Initially, S, a statistical surrogate was defined such that $R \perp\!\!\!\perp T|S$ (Prentice 1989). However, this conditional independence does not generally hold in the presence of confounders, (Lauritzen 2003). This has prompted alternative

definitions of surrogates. Consider the following example:

Example 5.2.5 *Viral Load as surrogate for disease status*

The viral load of an HIV sufferer can tell a doctor at what stage the disease is, and hence is a surrogate marker for the disease status. Together with other variables that describe the disease status, the viral load can help a doctor decide whether to initiate the next phase in treatment. DAG 5.5 represents the relationships between the initial treatment, the viral load as a surrogate marker, and the disease status D as a vector of disease status. The viral load variable V is then the basis for the decision to initiate the next phase of treatment.

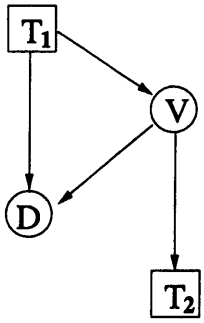

Figure 5.5: T_1 is the initial treatment, the viral load V is a surrogate for the status of the disease D and basis for the next stage of treatment T_2

If it is possible to estimate the magnitude of the relationships between T_1, V and D, from past patients, then it is possible for a doctor to have an idea of the disease status for a new patient based only on T_1 and V. Thus he or she will decide what treatment T_2 to assign next depending on what V is for the new patient.

5.3 Pearl's Controlled and Natural direct Effects

In his paper on direct and indirect effects, Pearl (2001b) tackles the problem of identifying the direct effect of a treatment using the counterfactual framework described in section 3.3.

Pearl describes two situations. First, he assumes that it is possible to intervene on C, the mediating variable, and he describes the conditions for *experimental* identification. This will be the focus of this chapter. In the second, he relaxes this assumption and describes conditions for *non experimental* identification.

In Pearl's framework, experimental identification refers to the circumstance where both the treatment and mediating variable are intervened upon. Non-experimental identification refers to the circumstance where neither the treatment nor the mediating variable are intervened upon.

Given that it is possible to intervene on the value of C, there are two ways in which, under certain circumstances, the direct effect of T on R can be identified.

The first is to *block* C at some value c and then change the treatment and record the response. Pearl calls this the *controlled* direct effect. The second way is as follows; Say the treatment is binary, with $T = 0$ or 1. Let T be set at 1; then the direct effect is calculated by keeping C at the value it *would have had* if T *had been* 0 and recording the value of R. This is called the

natural direct effect.

In order to clarify the concepts and definitions, we follow example 5.2.1.

Example 5.3.1 *Treatment with side effect continued*

Let the administration of the new drug treatment be a binary variable T. T takes on value t when the treatment is administered, and t^* when no treatment is administered. Let R be the response, also, let C be the mediating variable *taking aspirin.* For the sake of simplicity, C is binary, with values c if an aspirin is taken and c^* if it is not.

In Pearl's notation, R_{tc} is the response given $T = t$ and $C = c$. This variable would be counterfactual if the administered treatment was t^*. Finally C_t is the value of the mediating variable if $T = t$; as with the response this will also be a counterfactual variable if T is t^*. When referring to a specific unit u, (u) is added to the above variables. Thus the response for unit u is $R(u)$.

5.3.1 Controlled Direct Effect

In accordance with the counter factual framework, the direct effects are defined at the unit level first and then at the population level.

The **unit-level controlled direct effect** when C is blocked at c is defined for unit u as

$$CDE_c(t, t^*; R, u) = R_{tc}(u) - R_{t^*c}(u). \tag{5.1}$$

In terms of the above example this is the difference between the unit's re-

sponse to treatment $T = t$ and no treatment $T = t^*$ when in both cases an aspirin was not taken, that is $C = c$. One of the two variables is counterfactual.

The **average controlled direct effect** when C is blocked at c is the expectation over all u of (5.1);

$$CDE_c(t, t^*; R) = E(R_{tc} - R_{t^*c}). \tag{5.2}$$

Although (5.2) is in terms of counterfactual variables, it can be rewritten in non counterfactual terms in the augmented semantics covered in Chapter 2, with the additional conditional independence (implicit in the counterfactual framework) $R \perp\!\!\!\perp F_C | F_T, T, C$. Thus:

$$E(R|F_T = t, F_C = c) - E(R|F_T = t^*, F_C = c) \tag{5.3}$$

Both the effect of setting $T = t$ and of setting $T = t^*$ while holding the value of C fixed at c on R can in principle be calculated from experimental data where *both* T and C are intervened upon.

5.3.2 Natural Direct Effect

The **unit-level natural direct effect** is defined in terms of nested counterfactuals. The expression for a unit-level natural direct effect is:

$$NDE_{t^*}(t, t^*; R, u) = R_{t,C_{t^*}}(u) - R_{t^*}(u). \tag{5.4}$$

In words this is the difference between the effect of setting $T = t$ keeping C fixed at the value *it would have had* if, counter to fact, T had been set to t^*; and the effect on R of setting T to t^*.

Consider the patient John, who is receiving some treatment T. The natural direct effect for him is the difference between *a)* and *b)* below

a) John's disease status given that John was administered whichever drug treatment ($T = t$) and given that he was also administered the aspirin dose (c or c^*) he would have taken if he had not been given the drug treatment.

b) John's disease status if he had not been administered the drug treatment ($T = t^*$).

This means that it would be necessary to somehow find the John's *natural* aspirin intake outside the context of the disease and treatment scenario. Pearl assumes that this is a variable with a probability distribution. Further, for this to work, Pearl must assume that the disease itself is not related to headache occurrence. If it is, then it will confound the effect of the treatment and that of the aspirin.

As in the case of the controlled direct effect, the **average** natural direct effect is defined as follows:

$$E(R_{t,C_{t^*}}) - E(R_{t^*}). \tag{5.5}$$

If we look at this quantity, we see that the first part of (5.5) can be identified if we administer t to a group of patients and then administer the dose of aspirin they would each naturally have taken if they had not been administered the drug. In order to do this, we would also have to know what their natural

aspirin intake was, and also assume that the patient takes the same dose for every headache episode they experience. This is required because if we have a selection of different aspirin doses then we have more than one setting iof C, and different doses could result in different indirect effects. These details could have been asked in a survey.

The second part is always identifiable in principle as it is just the response to the baseline treatment.

Finally, Pearl defines the **Total effect** of T on R as the difference in expectations of R given $T = t$ and $T = t^*$.

$$E(R_t) - E(R_{t^*}). \tag{5.6}$$

Both parts of this equation are identifiable from experiment.

5.3.3 Experimental Identification

Given the quantities of interest as defined in the previous section, Pearl proceeds to determine formally, under what circumstances it is possible to identify them.

Theorem 5.3.1 (Experimental Identification) *Suppose there exists a set W of covariates such that $W \perp\!\!\!\perp T$ and such that*

$$R_{tc} \perp\!\!\!\perp C_{t^*} | W \quad \textit{for all } c, \ t \textit{ and } t^*. \tag{5.7}$$

This is represented in DAG 5.6. That is, the response R when T is set to t (i.e. $F_T = t$) and C is set to c is independent of C when C arises naturally

from its relationship with T when T is set to t^*, given the covariates W. When this is the case, then the average natural direct effect is experimentally identifiable, and is given by

$$\begin{aligned} NDE_{t^*}(t, t^*; R, u) &= E(R_{t,C_{t^*}}) - E(R_{t^*}) \\ &= \sum_{w,c} \Big([E(R_{tc}|w) - E(R_{t^*c}|w)] \\ &\quad \times P(C_{t^*} = c|w) P(w) \Big). \end{aligned} \tag{5.8}$$

Before looking in more detail at the assumptions other than conditional independence (5.7) that Pearl makes in order to estimate (5.8), let us take a quick look at (5.8): By conditional independence (5.7), the expectations after the summation can be expressed entirely in terms of R, F_T and W (formally shown in (5.11) later) and hence, given that these are all observed, can be identified from experimental data.

5.3.4 Assumptions underlying Theorem 5.3.1

We now look in some detail at the assumptions Pearl makes in order to ensure experimental identification. We reiterate that by experimental, Pearl means that intervention is possible both at T and at C.

There follows a discussion of *(i)* the relationship between $R_{tc}(u)$ and $C_{t^*}(u)$ the variables on the left hand side of the conditional independence symbol in (5.7), *(ii)* the role of W and how it can be interpreted in a real sense and finally, *(iii)* (5.8) is looked at in some detail.

$R_{tc}(u)$ and $C_{t^*}(u)$

Pearl assumes that the values of $R_{tc}(u)$ and $C_{t^*}(u)$ exist and are well-defined. This is in accordance to the counterfactual framework described in 3.3. In terms of John the patient, this means that at some level, both the variable that is his *response to treatment* $T = t$ and the variable that is his *headache status when he does not receive treatment* $T = t^*$ are independent conditional on the value of W. This is a concept that is hard to define as these quantities cannot be observed simultaneously at the individual level.

Further, there is no way of testing this independence on the individual level. On the population level, it would be possible to test such an independence, however, it is difficult to understand what the conditional independence means on a population level.

For the remainder of this section, when using R_{tc} and C_{t^*} we mean $R_{tc}(u)$ and $C_{t^*}(u)$.

What is W?

The existence of a set of covariates W, that are non-descendants of T or C and such that (5.7) holds is required for experimental identification in Pearl's set-up.

The DAG in figure 5.6 describes the relationship between T, R, C and W if W is a non-descendant of T or C or both, and further, if T is a decision node (i.e. it is always intervened upon).

What is the role of W? Is it possible for $W = \{\emptyset\}$? If so, what does this

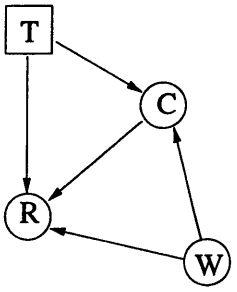

Figure 5.6: W is a non-descendant of T or C.

imply? What does it mean in terms of John and his treatment?

If W is empty, then $R_{tc} \perp\!\!\!\perp C_{t^*}$ marginally, that is, the response of R to treatment $T = t$ does not depend on the the value of C when $T = t^*$. Consider what this means in terms of John's treatments; Let us assume that John has a natural aspirin intake distribution. If we assume that $R_{tc} \perp\!\!\!\perp C_{t^*}$ marginally we are saying that

1. John's disease status when John is administered the drug treatment and takes aspirin as a consequence of the headache side-effect, and

2. John's natural aspirin intake

are independent. Making W empty might be reasonable in some cases, but does not seem to be so in this case as it is possible that these two variables are both related to a further variable that represents John's health status. It is therefore necessary to find a set of covariates that will make the two independent when conditioned upon. A set of general health status or life-style variables that relate to headache occurrence could be W.

However, it is not immediately clear how to choose W. Pearl chooses it according to whether it fulfils conditional independence (5.7) or not. This conditional independence does not have any meaning in the decision theoretic setting and it is thus unclear how it could be tested empirically. Further the choice of W might affect the value of (5.8), meaning that the natural direct effect is potentially not well-defined.

It is my opinion that Pearl must introduce W in order to make the causal effect (5.5) identifiable at least in principle.

Summary

By experimental Pearl means that both T and C can be intervened upon and the response to these interventions can be observed. In particular, the right hand side of (5.5), can be rewritten as

$$E(R_{t,C_{t^*}}) - E(R_{t^*}) = E(R_{t,C_{t^*}}) - E(R_{t^*,C_{t^*}}), \tag{5.9}$$

as the value of C when $T = t^*$ would indeed be C_{t^*}. This means that in this scenario, C is held constant at the value (or set of values) it takes on when the treatment is t^*.

Consider the first term in the first line of equation (5.8), this can be written as

$$E(R_{tC_{t^*}}) = \sum_w \sum_c \overbrace{E(R_{tc}|C_{t^*} = c, W = w)}^{(a)} P(C_{t^*} = c|W = w)P(W = w) \tag{5.10}$$

By conditional independence (5.7), *(a)* in (5.10) becomes $E(R_{tc}|W = w)$,

which is identifiable in principle by setting $T = t$ and $C = c$. In fact, each term after the summation in (5.8) is identifiable from experimental data in which both T and C are set to particular values.

It is therefore necessary to assume the following for identification of (5.8). from experiment in Pearl's framework:

1. W exists,

2. the conditional independence (5.7) holds

Equation (5.8) in non-counterfactual terms

Let us look at (5.8) under experimental conditions in terms of the decision theoretic notation developed so far. Let F_C denote the intervention variable on C. That is, F_C is such that when it is equal to a value $c \in \mathcal{C}$ the domain of C, then C is also equal to c, and when $F_C = \emptyset$ then C arises naturally from its relationship with $\mathbf{T}$.

Further note that $\mathbf{T}$ is a decision variable and hence has no parents. This is indicated by the bold face type.

The requirement that W be a nondescendant of $\mathbf{T}$ and C turns into the conditional independence $W \perp\!\!\!\perp (\mathbf{T}, F_C)$. There is a further conditional independence for R, which can be read off DAG 5.6, namely, $R \perp\!\!\!\perp F_C | \mathbf{T}, C, W$. Then:

$$\begin{aligned}(5.8) \quad = \quad & \sum_w \sum_c [\{E(R|\mathbf{T} = t, F_C = c, W = w) - E(R|\mathbf{T} = t^*, F_C = c, W = w)\} \\ & \times p(C = c|\mathbf{T} = t^*, F_C = \emptyset, W = w)\end{aligned}$$

$$\times p(W = w|\mathbf{T} = t^*, F_C = c)]. \tag{5.11}$$

This can be simplified to

$$\begin{aligned}(5.8) \quad = \quad & \sum_w \sum_c [\{E(R|\mathbf{T} = t, W = w, C = c, F_C = \emptyset) \\ & -E(R|\mathbf{T} = t^*, C = c, W = w, F_C = \emptyset)\} \\ & \times p(C = c|\mathbf{T} = t^*, F_C = \emptyset, W = w) \\ & \times p(W = w|F_C = \emptyset)]. \end{aligned} \tag{5.12}$$

If we then take $W = \emptyset$, then this reduces to

$$\sum_c \{E(R|\mathbf{T} = t, C = c, F_C = \emptyset) - E(R|\mathbf{T} = t^*, C = c, F_C = \emptyset)\} p(C = c|\mathbf{T} = t^*, F_C =$$

.

Essentially, in the non-counterfactual framework W can be seen as a stratifying variable. This has some interesting consequences. First, different choices of W will result in different direct and indirect effects. Thus, the direct effect for W representing a set of health-related variables minus sex will be different from W representing the same set of health-related variables with sex.

This again begs the question *What criteria are used to select it?* We shall see that, W is not fundamental for the development of the decision theoretic approach. In fact, it emerges that unless we have specific questions we want to ask where stratifying is important, or we suspect that there is a confounding variable, W is not a necessary element of the direct indirect effects framework in the decision theoretic model.

5.4 The non-counterfactual model for direct-indirect effects

As this thesis is based on a decision theoretic approach to causal inference, the concepts tackled by Pearl (2001b) must be re-expressed in non-counterfactual terms and defined formally. We have looked at (5.8) in non-counterfactual terms already. In this section, we construct the 3 variable direct indirect effects framework and define the effects in decision theoretic terms. It is a 3 variable framework as it is uniquely defined by the three observed variables $\mathbf{T}$, R and C, and their relationship to one another. In section 5.6, this will be extended to include another variable, uniquely defining the 4 variable direct indirect effects framework.

The key to the direct indirect effects framework developed in this chapter, is the manipulation of the mediating variable C via the point intervention node F_C and the more complex randomised manipulation variable M_C.

In the decision theoretic approach to this problem, T is always assumed to be intervened upon and thus there is no need to distinguish between F_T and T, making it simpler to represent the treatment as a single decision variable $\mathbf{T}$. Formally, $F_T \equiv \mathbf{T}$ and $F_T \neq \emptyset$ for the remainder of the chapter.

Before defining the regimes on C and the subsequent 3 variable direct indirect effects framework, it is necessary to define the elements involved.

5.4.1 Definition of Elements

The following are the elements that uniquely and completely define the decision theoretic direct indirect effects model for 3 variables.

List 5.4.1 Elements of 3 variable direct indirect effects model

1. A treatment variable $\mathbf{T}$, which is completely controlled by the experimenter. $\mathbf{T}$ takes on values in $\mathcal{T}$ which is the set of available treatments. $\mathbf{T}$ is in bold type to clarify that it is a decision variable, not a chance variable. For the sake of simplicity, $\mathbf{T}$ is binary, taking on value 1 if treatment took place and 0 otherwise.

2. A response variable R. This variable is always allowed to arise naturally. It takes on values r in $\mathcal{R}$.

3. The *mediating variable* C. This variable takes on values c in $\mathcal{C}$. It mediates part of the effect that $\mathbf{T}$ has on R.

4. F_C, the point intervention variable on C. This enables us to set C to a specific value.

5. M_C, the randomised intervention variable on C. In simple terms , it enables us to generate C from a specified distribution. It will be defined in detail in the next section. This variable is called M_C for *manipulation of* C.

6. The following conditional independence relationships

$$R \perp\!\!\!\perp F_C | C, \mathbf{T} \tag{5.13}$$

$$R \perp\!\!\!\perp M_C | C, F_C, \mathbf{T}. \tag{5.14}$$

The conditional independences are separated (they could be expressed in one conditional independence $R \perp\!\!\!\perp (F_C, M_C) | C, \mathbf{T}$) because the nature of the two intervention variables is different. F_C can be seen as representing a point intervention in the *real* world, whereas M_C is a tool that is used to clarify and represent potential randomised manipulations. This is defined in section 5.4.3. Note that the intervention and decision variables are all marginally independent of one another. This is trivial in so far as there is never any uncertainty associated with them. Also, the conditional independence (5.14) or (5.13), together with the constraint that $\mathbf{T}$ is a decision node and finally, the premise that there are no unobserved variables that are of interest, uniquely define the direct indirect effects DAG in 5.7. If we did not introduce any intervention variables on C then there would be DAGs that are Markov equivalent to the DAG in 5.7.

5.4.2 Regimes described informally

Although it may not physically possible to intervene on C, let us imagine nonetheless that we can intervene on C. What types of interventions would it be necessary to perform with the aim of estimating direct and indirect effects of treatment on response?

It is easiest to explain by referring directly to example 5.2.1. Setting ethical considerations aside, it is physically possible for an experimenter to control the administration of aspirin, which is the mediating variable in this example. Say that the experimenters are interested in 2 different dosages (none or 2 tablets) of aspirin. Let $C = \{0, 1\}$, for no aspirins and 2 respectively. There follows an informal description of the regimes:

5.4.2 Regimes

1. The pharmaceutical company is interested in the effect the aspirin has on the response. One randomly chosen group of patients is assigned level 0 of aspirin and randomly assigned the treatment, while a second group is assigned level 1, and randomly assigned the treatment. It is now possible to make inference about the direct effect of treatment for the two levels of aspirin intake by estimating $P(R|\mathbf{T}, C = c)$ for $c \in \{0, 1\}$. This is similar to what Pearl calls the controlled direct effect.

2. The pharmaceutical company has conducted a trial where one group of patients are not given the treatment and their *natural* aspirin intake when they do not receive treatment has been recorded and $p(C|\mathbf{T} = 0)$ estimated. A second group (exchangeable with the first) is run and again, the *natural* aspirin intake under treatment $\mathbf{T} = 1$ is observed. Hence $p(C|\mathbf{T} = 1)$ is also estimated. Then a new group of patients is recruited (exchangeable with the previous two groups), and the following regime is applied: Given that a patient is administered treatment

$\mathbf{T} = 0$, he is administered aspirin dosage according to the distribution of $C|\mathbf{T} = 1$, and vice versa for a patient administered $\mathbf{T} = 1$. The data gathered allows us to calculate the direct effect of treatment by comparing the responses to the two different treatments while generating the mediating variable from the same conditional distribution. This approach is similar to Pearl's natural direct effect. We need to assume that the three groups of patients are exchangeable before treatment is administered.

3. If it is not possible to observe the natural distribution of $C|T$, because running two trials is too expensive, but still possible to intervene on C, then the natural distribution can be replaced by a suitable distribution that generates values of C that are appropriate for the context. In the above example, the values of C could be generated by tossing a coin. It would then be possible to calculate the direct effect by simply summing out over C. One trial would still have to be run.

The last two approaches are analogous to direct and indirect standardisation. This aspect will be discussed in more detail after the formal definition of the regimes. It will not usually be the case that the results of the last two will be the same. However, they will provide comparable estimates of the effects.

5.4.3 Formal definition of Regimes on C

We proceed now to a formal definition of the regimes on C. These are all in terms of C's conditional distribution given its parents T, F_C and M_C as

seen in Figure 5.7. F_C and M_C are fundamentally different types of intervention variables as will become clear in the definition of the regimes, and later discussion in section subsection *M_C and F_C* in 5.4.3. This difference is reinforced in the DAGs where F_C has a standard decision node box and M_C has a rounded box.

First consider F_C; it has two types of settings, it is either idle $F_C = \emptyset$, or $F_C = c$ where $c \in \mathcal{C}$. When it is idle, C arises from its relationship with its other parents, when $F_C = c$, then C is *set* to the value c with no uncertainty.

M_C is context specific, meaning that it is uniquely defined by how it is related to C, T and R, that is in terms of the conditional independence relationship (5.14) and the regimes defined below. A change to any of these elements would require a redefinition of M_C.

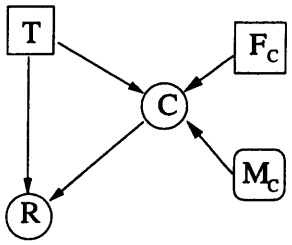

Figure 5.7: DAG describing the basic setup.

Regimes of C

The natural conditional distribution of $C|\mathbf{T} = t$ is denoted by P_t for $t \in \{0, 1\}$. The collection of conditional distributions P_t is denoted by $\mathcal{P}$.

Regime 5.4.1 *The first regime represents the observational case where C is allowed to arise naturally from its relationship with* $\mathbf{T}$. *This regime is observational with respect to C (not* $\mathbf{T}$*), and specifies the conditional probability distribution of C as P_t if and only if $\mathbf{T} = t$. This regime is indicated by*

$$F_C = M_C = \emptyset.$$

Thus the distribution of $C|\mathbf{T} = t, F_C = M_C = \emptyset$ is P_t.

Regime 5.4.2 *We say C is* ***set*** *if C has been forced to take on the value c^*. We indicate this by*

$$F_C \;=\; c^* \in \mathcal{C},$$

M_C can be anything as C is independent of M_C if it's value is set. We can therefore simply assume that $M_C = \emptyset$. Thus, the conditional distribution of $C|\mathbf{T} = t, F_C = c^*, M_C = \emptyset$ is δ_{c^*}, where δ_{c^*} is such that $P(C = c) = 1$ if $c = c^*$ and 0 otherwise.

The intervention results in an additional (if trivial) conditional independence;

$$C \perp\!\!\!\perp (T, M_C) | F_C \neq \emptyset,$$

with the corresponding change in the DAG as seen in figure 5.8. Note that when C is set to a fixed value c^*, the links from T to C and M_C are severed.

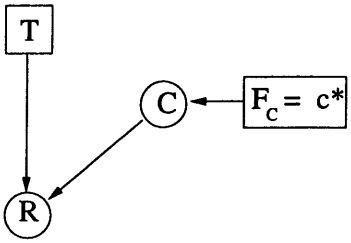

Figure 5.8: Setting $C = c^*$ via intervention node F_C. M_C is suppressed as it is idle.

Regime 5.4.3 *We next consider the case where C is sampled from conditional distribution P_{t^*}, $t^* \in \mathcal{T}$. This regime is indicated by*

$$\begin{aligned} F_C &= \emptyset, \\ M_C &= t^* \in \mathcal{T}. \end{aligned}$$

Hence, the conditional probability distribution of $C|\mathbf{T} = t, F_C = \emptyset, M_C = t^*$ is P_{t^*}. The associated graph is given in figure 5.9. As the value of T does not

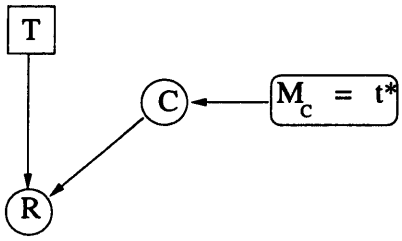

Figure 5.9: C is generated from the conditional distribution of $C|\mathbf{T} = t^*$

enter, we have the additional conditional independence:

$$C \perp\!\!\!\perp \mathbf{T} \mid M_C = t^*, F_C = \emptyset. \tag{5.15}$$

Finally, we introduce

Regime 5.4.4 *In this, C is generated from a specified distribution P_D. We indicate this by*

$$\begin{aligned} F_C &= \emptyset, \\ M_C &= D \in \mathcal{D}. \end{aligned}$$

So that the distribution of $C|\mathbf{T} = t, M_C = D$ is just D, that is C is sampled from $p(D = d)$, where the domain of C is the same as the domain of D, and $\mathcal{P}_\mathcal{D}$ is the set of suitable probability distributions. The associated graph and additional conditional independence are as with regime 5.4.3, where D replaces t^* and P_D replaces P_{t^*}, as the value of T does not enter.

M_C and F_C

The difference between F_C and M_C is made clear in the regime definition above. F_C can be defined independently of the regimes and any probability distribution whereas M_C cannot. M_C can no longer be used if an additional variable is added to the problem as discussed in section 5.6, F_C remains unchanged.

F_C is a point intervention and represents an intervention which is, in principle at least, possible, irrespective of any information we may have on the processes that govern the distribution of C. That is, we can in principle administer aspirin to a patient in a trial for $\mathbf{T}$, irrespective of a) whether we know if he has been administered treatment $\mathbf{T} = 1$ or $\mathbf{T} = 0$; b) whether we believe there is any relationship between the treatment and the occurrence of

headache or the intake of aspirin and the response, and c) whether we know the distribution of $C|\mathbf{T}$ or not.

M_C is fundamentally different as it represents a randomised intervention. Although physically possible, M_C is fictional in a sense that F_C is not. First, interventions of the type M_C represents are not normally preformed, it is simply a tool which allows us to express and identify the causal qualities we are interested in within the 3 variable framework. Second, whereas interventions represented by F_C can be performed with no knowledge (or assumptions) about how the underlying process works, $M_C = t^*$ can only be used once the assumptions have been made, the system observed, and inference made about the conditional distribution of C given $\mathbf{T}$.

$M_C = D$ *can* be performed without any knowledge of the workings of the system, in fact it is used when no knowledge about how C depends on $\mathbf{T}$ is available, and a generated direct effect is required. M_C in this incarnation has even less bearing on the *real* world than it has in $M_C = t^*$.

In light of the above discussion, it is clear why the conditional independences (5.13) and (5.14) are separated. One is the real part, $R \perp\!\!\!\perp F_C|\mathbf{T}, C$ in so far as it is testable, and the other the fictional part $R \perp\!\!\!\perp M_C|\mathbf{T}, F_C, C$, which is an assumption we make.

Regimes 5.4.3 and 5.4.4 as forms of standardisation

The last two regimes are similar to direct standardisation. The classic example of direct standardisation is that of mortality rates. We are interested in

comparing the mortality rates of two disjoint finite populations Ω and Δ. The probability of death in the two populations given by $p_\Omega(death)$ and $p_\Delta(death)$, do not offer a good comparison, as they do not reflect the age composition of the population. For instance, let Ω be the population of Eastbourne, a seaside town in England famous for its large population of pensioners, and Δ the population of the city of London. A straightforward comparison of $p_\Omega(death)$ and $p_\Delta(death)$ would not give the result we are really interested in as

$$p_\Omega(death) = \sum_{1 \leq k \leq K} p_\Omega(death|age = k)p_\Omega(age = k),$$

and the age composition $p_\Omega(age = k)$ for $k = 1, \ldots, K$ is not necessarily the same as $p_\Delta(age = k)$. What we care about is a pairwise comparison of the age specific death rates, and we may not have the age distributions of the population. A solution is to introduce a standard age composition $p^S(age = k)$ for each $k = 1, \ldots, K$, and use this to find a standardised death rate for the two populations which are then comparable.

$$p^S_\Omega(death) = \sum_{1 \leq k \leq K} p_\Omega(death|age = k)p^S(age = k)$$

A similar argument for Δ gives a standardised rate of $p^S_\Delta(death)$.

As the choice of $p^S(age = k)$ for $k = 1, \ldots, K$, is arbitrary, we can always standardise using the age composition of one of the populations, say $p_\Omega(age = k)$. This is in fact analogous to what regime 5.4.3 is doing. It is standardising the effect of treatment on response using the conditional distribution of $C|\mathbf{T} =$

t^* the baseline value of $\mathbf{T}$. Regime 5.4.4 is more akin to using a standard or historical age composition.

5.4.4 Definition of the 3 variable d-i effects model

Now define the *3 variable d-i effects model* as $\Delta = \{\mathbf{T}, R, C, F_C M_C, \mathcal{I}, \mathcal{P}, \mathcal{P}_{\mathcal{D}}\}$. This collection completely determines the model; the treatment $\mathbf{T}$, response R, the mediating variable C, the intervention variables F_C and M_C of C, with their corresponding regimes and the conditional independence $\mathcal{I} = \{R \perp\!\!\!\perp (F_C, M_C) | \mathbf{T}, C\}$ and the collections of probability distributions $\mathcal{P}$ and $\mathcal{P}_{\mathcal{D}}$.

Definition of Direct effects in the 3 variable d-i effects model

We can now define decision theoretic versions of the average controlled direct effect (5.2), and of the natural direct effect (5.5) in terms of F_C and M_C.

Definition 5.4.1 (Direct effect for C set at c) *The direct effect of* $\mathbf{T} = t$ *with respect to baseline* $\mathbf{T} = t^*$ *on response* R *for* C *set at* c *is given by*

$$E(R|\mathbf{T} = t, F_C = c, M_C = \emptyset) - E(R|\mathbf{T} = t^*, F_C = c, M_C = \emptyset). \quad (5.16)$$

Call this effect the SE_c *(the Set Effect).*

Definition 5.4.2 (Direct effect for C generated from P_{t^*}) *The direct effect of* $\mathbf{T} = t$ *with respect to the baseline* $\mathbf{T} = t^*$ *on response for* C *generated from* P_{t^*} *the distribution of* C *conditional on* $\mathbf{T} = t^*$ *the baseline treatment is*

given by

$$E(R|\mathbf{T}=t, M_C=t^*, F_C=\emptyset) - E(R|\mathbf{T}=t^*, M_C=t^*, F_C=\emptyset)$$
$$\equiv E(R|\mathbf{T}=t, M_C=t^*, F_C=\emptyset) - E(R|\mathbf{T}=t^*, M_C=\emptyset, F_C=\emptyset). \quad (5.17)$$

Call this effect the GDE_t (the Generated Direct Effect from P_t).

In this framework, it is possible to go one step further and define another direct effect, based on regime 5.4.4.

Definition 5.4.3 (Direct effect for C generated from D) *The direct effect of $\mathbf{T}=t$ with respect to the baseline $\mathbf{T}=t^*$ on response for C generated from a specified and appropriate distribution D.*

$$E(R|\mathbf{T}=t, M_C=D, F_C=\emptyset) - E(R|\mathbf{T}=t^*, M_C=D, F_C=\emptyset) \quad (5.18)$$

Call this effect the GDE_D (the Generated Direct Effect from D.)

Definition 5.4.4 (Total effect) *The total effect of $\mathbf{T}=t$ with respect to the baseline $\mathbf{T}=t^*$ on response is given by*

$$E(R|\mathbf{T}=t) - E(R|\mathbf{T}=t^*). \quad (5.19)$$

This effect is referred to as the TE.

Note that it is also possible to look at the GDE_{t^*}, that is

$$E(R|\mathbf{T}=t^*, M_C=t, F_C=\emptyset) - E(R|\mathbf{T}=t, M_C=\emptyset, F_C=\emptyset),$$

where we take the treatment $\mathbf{T} = t$ as the baseline.

Finally, the indirect effect can be defined as the difference between the total effect and the generated direct effect. As there are two types of generated direct effects, one resulting from GDE_{t^*} and another from GDE_D.

Definition 5.4.5 (Indirect effect) *The indirect effect when C is sampled from $C|\mathbf{T} = t^*$ is*

$$IE_{t^*} = TE - GDE_{t^*}. \tag{5.20}$$

The indirect effect when C is sampled from D is given by

$$IE_D = TE - GDE_D. \tag{5.21}$$

Direct effects expressed without M_C

Consider briefly equation (5.17), where we define the direct effect when C is sampled from P_t. It can be expressed without M_C as follows:

$$\begin{aligned}\sum_c \{&E(R|\mathbf{T} = t, C = c, F_C = \emptyset) \times p(C = c|, T = t^*, F_C = \emptyset) \\ &+E(R|\mathbf{T} = t, C = c, F_C = \emptyset) \times p(C = c|, T = t, F_C = \emptyset)\}.\end{aligned} \tag{5.22}$$

See the next section for the process of turning this equation into equation (5.17). In this definition, as in that of the GDE_D, M_C serves to simplify and clarify what is otherwise an unwieldy mix of probabilities. The *fictional* nature of M_C has already been discussed in section 5.4.3, and the above reinforces the fact that it is not an essential part of the definition of direct effects.

Direct effects as standardisation

Expression (5.22) of the direct effect generated from P_{t^*} also clearly shows the relationship between standardisation and direct effects formulation. In particular, equation (5.22) is a case of *natural standardisation*, that is, the C is sampled from its distribution given $\mathbf{T}$, which is based on real data. If we replace $p(C = c|, T = t^*, F_C = \emptyset)$ with $p(D = c)$, in (5.22) above, we get the equivalent expression for GDE_D. We are now standardising *arbitrarily*, that is, we are sampling C from a distribution that does not necessarily have any bearing on what C is or how it is related to its ancestors.

5.5 Identification using the non-counterfactual model

Consider the above model, with C discrete. When is it possible to identify the effects under either data that is experimental with respect to C, that is where C is intervened upon directly, or observational data where C arises naturally from its relationship with $\mathbf{T}$? Recall that is it assumed that $\mathbf{T}$ is a decision variable and hence always intervened upon.

Consider first the direct effect for C set at c given by (5.16). It is clear that this effect can be estimated directly from experimental data where it is possible to set $C = c$. However, it can only be estimated from observational data if we can observe

$p(R|\mathbf{T} = t, C = c, M_C = \emptyset, F_C = \emptyset)$

and

$p(R|\mathbf{T} = t^*, C = c, M_C = \emptyset, F_C = \emptyset)$,

as $E(R|\mathbf{T} = t, F_C = c, M_C = \emptyset) \equiv E(R|\mathbf{T} = t, F_C = c, C = c, M_C = \emptyset) \equiv E(R|\mathbf{T} = t, C = c, M_C = \emptyset, F_C = \emptyset)$ by conditional independence (5.14).

Consider now the direct effect for C generated from P_{t^*}, (5.17), for the sake of simplicity, as $F_C = \emptyset$ for all of the effects below, it will be omitted. The second part can be estimated from data when $\mathbf{T} = t^*$. The first part of (5.17) is not so straightforward;

$$\begin{aligned} E(R|\mathbf{T} = t, M_C = t^*) &= \sum_c E(R|\mathbf{T} = t, M_C = t^*, C = c) \\ &\quad \times p(C = c|\mathbf{T} = t, M_C = t^*) \qquad (5.23) \\ &= \sum_c E(R|\mathbf{T} = t, C = c, M_C = \emptyset) \\ &\quad \times p(C = c|, M_C = t^*) \qquad (5.24) \\ &= \sum_c E(R|\mathbf{T} = t, C = c, M_C = \emptyset) \\ &\quad \times p(C = c|, T = t^*, M_C = \emptyset). \qquad (5.25) \end{aligned}$$

We go from (5.23) to (5.24) because $R \perp\!\!\!\perp M_C|T, C$ and $C \perp\!\!\!\perp T|M_C \neq \emptyset$. From (5.24) to (5.25) because $M_C = t^*$ is defined as drawing C randomly from the its conditional distribution given $T = t^*$.

The quantities in (5.25) are all identifiable from data where only $\mathbf{T}$ has been intervened upon when this is available on all three variables. That is, if we can completely observe $p(R|T = t, C = c, M_C = \emptyset)$ and $p(C|T = t^*, M_C = \emptyset)$.

Finally, the total effect can be identified from data where only $\mathbf{T}$ has been

intervened upon, which we are assuming is always the case.

Note that it is not necessary to introduce W in order to be able to identify these quantities in the non-counterfactual framework.

5.5.1 Identification using the GDE_D

Consider the generated direct effect from D given by (5.18). Again we drop the F_C for convenience. When would we be interested in using this effect? We may be unable to observe the true conditional distribution of $C|\mathbf{T}$, and therefore have to sample C from another distribution. D either be a distribution that is in some way related to the actual distribution of $C|\mathbf{T}$, or a distribution of another random variable with the same domain as C.

For example, in the treatment side-effect example, if we cannot actually observe the distribution of aspirin intake of the patients that are being treated, we may nonetheless have data on a previous study from which we can estimate the occurrence of headache.

Alternatively, we may have no empirical evidence to go on. We can choose a distribution for $C|\mathbf{T}$ based on our beliefs or just arbitrarily. For example, we may toss a coin if there are two possible aspirin dosages.

What is important is that the effect we find can still be used for inference as this type of *standardisation* allows us to compare the expectations in the effects.

Consider the first part of (5.18) dropping the F_C:

$$E(R|\mathbf{T}=t, M_C=D) \;=\; \sum_c E(R|\mathbf{T}=t, C=c, M_C=D)p(C=c|M_C=D)$$

$$= \sum_c E(R|\mathbf{T} = t, C = c, M_C = \emptyset)p(C = c|D),$$

where $p(C = c|D)$ simply means that C is being sampled from the distribution of D.

So, if it is possible to observe the response as well as $\mathbf{T}$ and C, then it is possible to identify this quantity even if we cannot observe the distribution of $C|\mathbf{T}$.

5.5.2 Identification when C is binary

Consider the simple scenario when C is a binary variable taking on values 0 and 1. In the running example, these could represent two aspirin dosages, no aspirin and one aspirin respectively.

Assume that it is possible to observe all the variables under natural conditions, and hence estimate the conditional distributions of C given $\mathbf{T}$ and R given C and $\mathbf{T}$.

That means that we can estimate

1. $p(R|\mathbf{T}, C = 0, F_C = \emptyset, M_C = \emptyset)$,
2. $p(R|\mathbf{T}, C = 1, F_C = \emptyset, M_C = \emptyset)$,
3. $p(C = 1|\mathbf{T}, F_C = \emptyset, M_C = \emptyset)$,
4. $p(C = 0|\mathbf{T}, F_C = \emptyset, M_C = \emptyset)$,
5. $p(R|\mathbf{T}, F_C = \emptyset, M_C = \emptyset)$,

for $\mathbf{T} = t$ and t^*.

The SE_0 and SE_1 can then both be identified from the above data as $p(R|\mathbf{T} = t, F_C = 1) \equiv p(R|\mathbf{T} = t, C = 1, F_C = 1) \equiv p(R|\mathbf{T} = t, C = 1, F_C = \emptyset)$. Similarly for $T = t^*$ and $C = 0$. We suppress M_C as is it idle. The TE is also immediately identifiable from $p(R|\mathbf{T})$.

The GDE_t can also be identified directly as $E(R|\mathbf{T} = t^*)$ can be estimated from the data and by (5.25),

$$\begin{aligned} E(R|\mathbf{T} = t, M_C = t^*) &= E(R|\mathbf{T} = t, C = 1, M_C = \emptyset)p(C = 1|, T = t^*, M_C = \emptyset) \\ &\quad + E(R|\mathbf{T} = t, C = 0, M_C = \emptyset)p(C = 0|, T = t^*, M_C = \emptyset). \end{aligned}$$

Where we suppress F_C as it is idle.

If it is possible only to estimate items 1,2 and 5 from the above list, that is, it is not possible to estimate the conditional distribution of C given T directly from the data, we can still estimate it indirectly.

Let $p = p(C = 1|\mathbf{T} = t^*, M_C = \emptyset)$, then $1 - p = p(C = 0|\mathbf{T} = t^*, M_C = \emptyset)$. Also, let $p_0 = p(R|\mathbf{T} = t^*, C = 0, M_C = \emptyset)$ and $p_1 = p(R|\mathbf{T} = t^*, C = 1, M_C = \emptyset)$. Then

$$\begin{aligned} p(R|\mathbf{T} = t^*) &= p_1 p + p_0(1 - p) \\ &= p_0 + p(p_1 - p_0). \end{aligned}$$

So

$$p = \frac{p(R|\mathbf{T} = t^*) - p_0}{p_1 - p_0}.$$

This formula can be plugged in as required in the effects equations.

5.6 Extensions

There are contexts in which simply considering the 3 variables in the framework developed so far is not sufficient. This happens when there is a variable that influences both the mediating variable and the response variable, such that the conditional independence (5.14) can no longer be assumed to hold or we are interested in the effects in different strata of the population.

We introduce one additional variable W, such that in addition to the existing conditional independences, $W \perp\!\!\!\perp T$ also holds. In graphical terms means this that it is a parent of both R and C but not of $\mathbf{T}$. As $\mathbf{T}$ is a decision node and therefore has no parents, this implies that W is not a child or a parent of $\mathbf{T}$.

W can be seen as a stratifying variable. This is demonstrated clearly in the example 5.6.1 below. It is often the case that if we do not stratify and estimate only population average quantities, we may base future decision making on inadequate results, as responses in different strata diverge. W can also be a type of confounder as seen in example 5.6.2 below. Finally, W can also be of interest if it can be intervened upon as shown in the example 5.6.3. As W is assumed to influence both the mediating variable and the response, it can be used to change the value or nature of the indirect and therefore total effect of treatment on response.

The setting regime and the point intervention variable are not considered for the extended case. This is because the setting regime simply severs the

links between the mediating variable and all other variables in the problem, and W in particular.

5.6.1 Examples motivating the introduction of W

Example 5.6.1 *W as sex*

Consider extending example 5.2.1. We may have reason to believe that the side-effect is worse for women than men, and that therefore they take more aspirin, resulting in a different mediated effect. This makes it necessary to separately evaluate the effects for men and women. We therefore add a variable W as exhibited in DAG 5.10 which represents sex, and evaluate the direct and indirect effects for each of the two settings of W instead of an average effect for the general population as has been considered so far.

Example 5.6.2 *W as Doctor indicator*

Another extension of the above example, where the confounding property of W is more apparent is as follows. The patients are administered a treatment by one of two doctors. One doctor is aware of the headache side-effect, and tells his patients not to take any pain killers. The other doctor is not aware of the problem, and tells his patients nothing. W indicates which of the two doctors treats a particular patient. If we do not know which patients are treated by which doctor, then our effects will be confounded.

Example 5.6.3 *W as education*

Consider example 5.2.4, where the introduction of costsharing is thought to

have a negative direct and indirect effect on the use of preventive services via GP check-ups. A sub-study is conducted which looks at the effect of introducing cost-sharing on women between 30 and 60. It emerges that women with cases of cancer in the family, tend to be more informed and regularly check themselves. These tend to go to both the GP and/or seek out preventive services more often that women without cases of cancer in the family, regardless of the cost. A variable W that represents the level of breast-cancer specific knowledge is introduced to the problem.

Although in an observational context, this knowledge occurs at random and is the consequence of family cancer history, it is of interest to the insurance company as levels of cancer specific knowledge can be intervened upon, say by sending health insurance policy holders information leaflets. This may counteract the negative trend costsharing has on the use of preventive services. In fact an intervention such as the sending of leaflets, represented as an intervention node, say F_W, is independent of the response given W, and hence the effect of a point intervention can be identified from the observational study on cancer in the family.

5.6.2 Definition of M_C^W

In order to include a new variable W, we define a new direct indirect effects framework in the presence of W. First, the elements are defined. The first three are the same as those defined in list 5.4.1. The following are appended after 3.

4. W, a chance variable that is a parent of both R and C.

5. M_C^W a manipulation variable on C.

6. The following conditional independences:

$$W \perp\!\!\!\perp (\mathbf{T}, M_C^W) \tag{5.26}$$

$$R \perp\!\!\!\perp M_C^W | \mathbf{T}, C, W. \tag{5.27}$$

It is worth looking at the graph that represents these conditional independences uniquely. This is given in figure 5.10

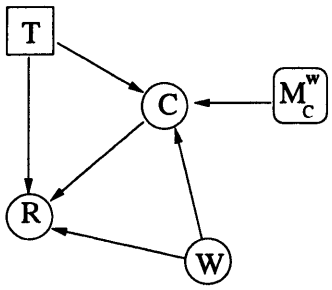

Figure 5.10: The extended framework

5.6.3 Regimes of M_C^W

The regimes on C in the presence of W, are only interesting to us when the relationship between C and W is preserved. If the relationship is severed, as in the setting regime or in the case where we sample C from P_{t^*} as in regime 4.3 in the 3 variable framework, then W is no longer important as it does not add anything to our knowledge of the direct and indirect effects. For

this reason, we do not consider the setting regime or the regime where C is generated from its conditional distribution given only $\mathbf{T}$.

Denote the conditional distribution of $C|\mathbf{T} = t, W = w$ by P_{tw} for $t \in \{0,1\}$ and $w \in \mathcal{W}$, where $\mathcal{W}$ is the domain of W. Further, denote the collection of P_{tw}s by $\mathcal{Q}$. As before $\mathcal{P}$ is the collection of conditional distributions of $C|\mathbf{T} = t$, P_t.

Regime 5.6.1 *The idle regime of M_C^W corresponds to the observational case on C, where it is allowed to arise naturally from its relationship with $\mathbf{T}$ and W. We indicate this regime by*

$$M_C^W = \emptyset,$$

so the distribution of $C|\mathbf{T} = t, W = w, M_C^W = \emptyset$ is P_{tw}

Regime 5.6.2 *Now consider another generating regime where C is sampled from the conditional distribution P_{t^*w}, that is from $p(C|\mathbf{T} = t^*, W = w)$, where $W = w$ is the **actual** value of W. This is indicated by*

$$M_C^W = t^*.$$

(Not to be confused with $M_C = t^*$, in the 3 variable framework.) This regime induces the conditional independence $C \perp\!\!\!\perp \mathbf{T}|M_C^W = t^*$, that is we are severing the link between $\mathbf{T}$ and C but not the link between C and W. Note that we see W and then sample from the appropriate P_{t^*w}.

This regime is represented by the DAG in 5.11.

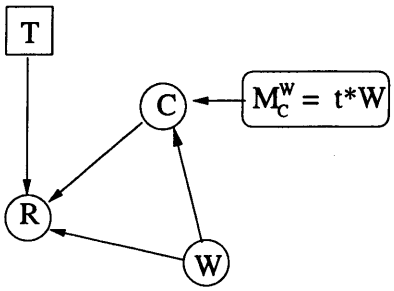

Figure 5.11: Regime 5.6.2

5.6.4 Definition of Effects using M_C^W

Now that the regimes have been defined, we can define the *4 variable d-i effect* as follows; $\Delta_4 = \{\mathbf{T}, R, C, WM_C^W, \mathcal{J}, \mathcal{Q}\}$. This collection completely determines the model, the treatment $\mathbf{T}$, response R, the mediating variable C, the confounding variable W, the regimes M_C^W of C, and the conditional independences assumed on the introduction of M_C^W, in this case $\mathcal{J} = \{R \perp\!\!\!\perp M_C^W | \mathbf{T}, C, W\}, \{W \perp\!\!\!\perp \mathbf{T}, M_C^W\}$ and finally the collection of probability distributions P_{tw}, $\mathcal{Q}$.

5.6.5 Using M_C^W

There are three general scenarios in which we will be interested in using M_C^W instead of M_C. These correspond to the three examples in section 5.6.1. In the first case we estimate effects conditional on specific values of W, in the second, we want to somehow eliminate W as it is a nuisance variable, and estimate overall direct effect. In the last example we are interested in manipulating W and need to determine whether it is possible to make causal

inference about the manipulation of W from data where W arises naturally.

W as a stratifying variable

In example 5.6.1, we are interested in estimating the direct and indirect effect for each sex. Let $W = 0$ when the patient is a man and $W = 1$ when the patient is a woman. The woman-specific direct effect of response on treatment is given by:

$$
\begin{aligned}
& E(R|\mathbf{T} = t, M_C^W = t^*, W = 1) - E(R|\mathbf{T} = t^*, M_C^W = t^*, W = 1) \quad (5.28) \\
&= \sum_c [E(R|\mathbf{T} = t, M_C^W = t^*, W = 1, C = c) \\
&\quad \times p(C = c|\mathbf{T} = t, M_C^W = t^*, W = 1) \\
&\quad - E(R|\mathbf{T} = t^*, M_C^W = t^*, W = 1, C = c) \\
&\quad \times p(C = c|\mathbf{T} = t, M_C^W = t^*, W = 1)] \\
&= \sum_c [E(R|\mathbf{T} = t, W = 1, C = c, M_C^W = \emptyset) \\
&\quad \times p(C = c|\mathbf{T} = t^*, W = 1, M_C^W = \emptyset) \\
&\quad - E(R|\mathbf{T} = t^*, W = 1, C = c, M_C^W = \emptyset) \\
&\quad \times p(C = c|\mathbf{T} = t^*, W = 1, M_C^W = \emptyset)] \\
= \sum_c [E(R|\mathbf{T} = t, W = 1, C = c, M_C^W = \emptyset) & p(C = c|\mathbf{T} = t^*, W = 1, M_C^W = \emptyset)] \\
&\quad - E(R|\mathbf{T} = t^*, W = 1, M_C^W = \emptyset)
\end{aligned}
$$

Call (5.28) the *generated direct effect conditional on* W as we sample C from the distribution P_{t^*w}, and then restrict our attention to the case where $W = 1$. The above is essentially equivalent to the 3 variable set-up conditional on the

value $W = 1$. In order to estimate (5.28), it is necessary to have fully observed the natural distributions of all the variables ($M_C^W = \emptyset$).

Again, it is clear that M_C^W is only a tool that permits us to express in more simple terms concepts and probabilities that are otherwise as seen above inelegant. Further, the role of the sampled probability of C as a standardising element is reinforced.

In a more general scenario, we are interested in estimating the direct and indirect effect within specific strata of W, and the quantity we are looking for is the direct effect conditional on $W = w$,

$$E(R|\mathbf{T} = t, M_C^W = t^*, W = w) - E(R|\mathbf{T} = t^*, M_C^W = t^*, W = w). \quad (5.29)$$

W as a confounder

In example 5.6.2, W is a nuisance variable in so far as it confounds the relationship between the mediating variable and the response variable. We are not interested in W for its own sake, but are interested in eliminating it. The quantity of interest is the direct effect of $\mathbf{T}$ on R via C;

$$E(R|\mathbf{T} = t, M_C^W = t^*) - E(R|\mathbf{T} = t^*, M_C^W = t^*) \quad (5.30)$$

Consider the first part:

$$\begin{aligned} E(R|\mathbf{T} = t, M_C^W = t^*) &= \sum_w \sum_c E(R|\mathbf{T} = t, M_C^W = t^* W, W = w, C = c) && (5.31) \\ &\quad \times p(C = c|\mathbf{T} = t, M_C^W = t^*, W = w) && (5.32) \\ &\quad \times p(W = w|\mathbf{T} = t, M_C^W = t^*) && (5.33) \end{aligned}$$

$$= \sum_w \sum_c E(R|\mathbf{T}=t, W=w, C=c, M_C^W=\emptyset) \quad (5.34)$$

$$\times p(C=c|\mathbf{T}=t^*, W=w, M_C^W=\emptyset) \quad (5.35)$$

$$\times p(W=w|M_C^W=\emptyset) \quad (5.36)$$

We go from (5.31) to (5.34) by conditional independence (5.27). From (5.32) to (5.35) as $C \perp\!\!\!\perp \mathbf{T}|M_C^W = t^*$ but it still depends on the actual value of $W = w$, and is therefore sampled from P_{t^*w}. Finally from (5.33) to (5.36) by conditional independence (5.26).

The second part of (5.30) is simply

$$E(R|\mathbf{T}=t^*, M_C^W=\emptyset), \quad (5.37)$$

where we sum out the values of W or ignore them. It is fundamentally the idle regime of M_C^W for $\mathbf{T} = t^*$. Again, it is necessary to have observed all the variables under natural conditions before we can make this type of inference. This might not be sensible in example 5.6.2 as we might not know what the doctor is saying to his patients. However, the quantity of interest, which is $p(C = c|T, W)$ can be observed.

Introducing F_W

As in example 5.6.3, we may be interested in a variable like W because it permits us to intervene on the indirect effect, and hence the total effect. If it is possible to intervene directly, then W can come to mirror $\mathbf{T}$.

The insurance company wants to know what effect intervening will have, before they actually try to do so. In other words, they want to be able to

estimate

$$E(R|\mathbf{T}=t, M_C^W=t^*, F_W=w) - E(R|\mathbf{T}=t^*, M_C^W=t^*, F_W=w), \quad (5.38)$$

where F_W is the point intervention node on W in the augmented notation. So, when $F_W = w$ where $w \in \mathcal{W}$, then $W = w$, and when $F_W = \emptyset$ then W arises naturally. Consider the graphical representation of the problem including F_W in figure 5.12. From it, we can see that F_W is independent of

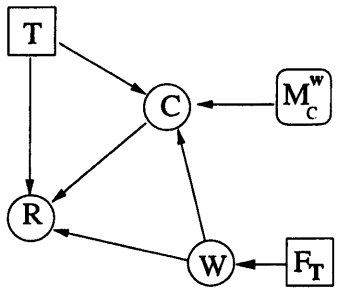

Figure 5.12: Introducing F_W

all other variables given W. Hence, (5.38) can be rewritten as

$$\begin{aligned} & E(R|\mathbf{T}=t, M_C^W=t^*, F_W=w, W=w) \\ & \qquad - E(R|\mathbf{T}=t^*, M_C^W=t^*, F_W=w, W=w) \\ = \; & E(R|\mathbf{T}=t, M_C^W=t^*, W=w, F_W=\emptyset) \\ & \qquad - E(R|\mathbf{T}=t^*, M_C^W=t^*, W=w, F_W=\emptyset) \quad (5.39) \end{aligned}$$

It is hence possible to estimate the direct effect of intervening on W from data where W arises naturally.

This is very useful for the insurance company, as it is possible to find out whether educating more women about cancer might encourage more of them

to seek preventive services and/or visit the GP, given that it is education that makes women with cancer more likely to check themselves.

Notice that we find ourselves in the same mathematical scenario as in the first example, where W is a stratifying variable. However, although the mathematical expression we want to estimate is the same, as (5.29) is equal to (5.39), the contexts out of which the two arise are very different. In the first example W is sex, a variable that cannot be intervened upon, in this example, W is the level of education about cancer, which is a variable that can at least in principle be intervened upon.

Identification

As in the case of the 3 variable framework, provided that each variable is observable, then the direct effect is in fact identifiable from observational data, that is when $M_C^W = \emptyset$.

5.7 Conclusions

It is easy to see how the framework can be extended if we believe that the 3 variables in the 3-variable framework are embedded in a larger network. It is possible to consider more than one mediating variable, either by making C a vector, or by creating manipulation variables for each mediator of interest. In this case, extra care will have to be taken if the mediators are related to one another.

We do not look at non-experimental identification in the way Pearl sees

it as we assume that the treatment variable is always intervened upon. Thus the problem is always experimental with respect to the treatment. This is because generally, when dealing with this problem, as seen in the examples (5.2), the data is experimental with respect to the treatment variable.

Also, by introducing the manipulation variable, we see that experimental and non-experimental identification with respect to the mediating variable are the same in the decision theoretic framework, given that we assume that the treatment is always intervened upon. And further, that identification is possible from purely observational data.

Again we see that if we make sensible assumptions, that is, the treatment variable is always intervened upon, it is possible to make casual inference on direct and indirect effects in the decision theoretic framework . That is, we do not need to make assumptions that are hard to understand or test such as conditional independence (5.7), or introduce additional variables the choice of which is not clear and changes the effects we are interested in such as W.

Chapter 6

Effect of Treatment on the Treated

6.1 Introduction

Consider the following story; Doctor A, working at hospital One is given a new drug treatment to administer to his patients. He chooses who to treat on the basis of a set of characteristics that are known only to him. The responses of those who were treated are recorded. The data is later given to a statistician to analyse. The information at her disposal are the treatment responses of those who were treated and whatever covariate information is in their medical records. The criteria used by the doctor to decide who to treat remain unknown.

The quantity of interest is the causal effect of treatment on the outcome, say disease status. However, there are problems with determining what this is. The selection criteria for treatment are unknown and act as confounders on the treatment effect.

An alternative solution to this problem was first proposed by Heckman and Robb (1985) . The idea is to estimate not the causal effect of treatment directly as we would like, but to estimate the *effect of treatment on the treated* (ETT). In a counterfactual/potential outcomes framework this is given by

$$E(Y_1 - Y_0|X, T = 1). \tag{6.1}$$

T is the treatment variable, Y_1 and Y_0 are the potential responses to treatment $T = 1$ and $T = 0$ respectively, and X is a set of known covariates.

A problem emerges when we look more closely at the ETT. Say that two doctors were given the new drug treatment to administer to the same patients. One doctor decides to treat her patients at random, while the other chooses to treat those who in his opinion would most benefit from the drug. Would the two doctors' data result in different values for the ETT? What quantities would have to be the same in order for the two doctors to have the same value for the ETT?

Another problem is that of identification (see Chapter 2 section 2.3.3). What type of data, observational, experimental or a combination of both is necessary in order to identify the ETT?

The first part of this chapter looks at the ETT from both the counterfactual as well as the decision theoretic point of view and we prove that it is a well-defined quantity. In the second part, we show that the ETT is identifiable from a combination of experimental and observational data, but that in order to identify it in practice from observational data, it is neces-

sary to make some strong assumptions. Some of the possible assumptions made in the counterfactual framework are discussed and reconsidered in a non-counterfactual framework.

6.2 Decision-theoretic setup

The doctor's story can be formalised as follows in *non-counterfactual* terms. Let T be the treatment, F_T its intervention variable and U the doctor's unknown selection criteria. We first assume the following;

$$U \perp\!\!\!\perp F_T. \tag{6.2}$$

This conditional independence is interpreted as follows; the criteria U, that the doctor uses to decide what treatment he would administer do not depend on how the treatment is administered under intervention. Denote by Y be response to treatment. We further assume that

$$Y \perp\!\!\!\perp F_T|(U,T). \tag{6.3}$$

In words, we believe that the response to the treatment does not depend on how the treatment was administered (in an experimental context or not), given the treatment was administered and the selection criteria known.

For conditional independence (6.3) to be reasonably assumed to hold it should generally include more than just the unknown selection criterion. This is because it is not generally the case that Y will be independent of F_T given T and unknown selection criteria. Usually, U will have to contain additional

information such as personal details, age and sex or general health status variables. For the sake of simplicity these will not be considered. If U is observable (which in the context we are discussing it typically is not), then the conditional independences (6.2) and (6.3) define it as a *sufficient covariate* (Lauritzen 2001)

The setup described by conditional independences (6.2) and (6.3) is expressed graphically in Figure 6.1.

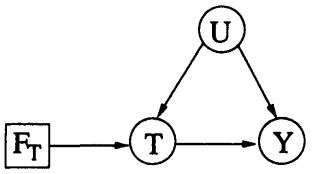

Figure 6.1: U is a potential confounder in the problem.

Let treatment T be a binary variable taking on values 1 for treatment, 0 for control. The effect of the treatment on the treated is defined in (6.4) below.

$$E[ACE(U)|T = 1, F_T = \emptyset] \tag{6.4}$$

where

$$ACE(U) =: E(Y|F_T = 1, U) - E(Y|F_T = 0, U). \tag{6.5}$$

Equation (6.4) is the expectation of the average causal effect given selection criteria U and treatment $T = 1$ having taken place. That is, the effect of treatment on the treated (ETT). This definition of it emphasises the fact that the ETT depends on the choice of U.

If we expand (6.4) we get (6.6) below:

$$E[\ E(Y|U, F_T = 1) - E(Y|U, F_T = 0)\ |T = 1, F_T = \emptyset]. \qquad (6.6)$$

This is the fully expanded form of the ETT. The inner expectations are over Y and the outer expectation is over U. This is because U is unobservable but Y is conditioned on it.

Note that if U and T were independent given $F_T = \emptyset$, for example if T were randomised, then (6.6) would simply reduce to $E(Y|F_T = 1) - E(Y|F_T = 0)$, the average causal effect. However, we cannot replace the conditioning on $T = 1$ in the outer brackets with conditioning on $F_T = 1$ as $U \not\perp\!\!\!\perp F_T|T$, although this can be done in the inner bracket.

Note that (6.6) can be rewritten as follows:

$$E(Y|T = 1, F_T = \emptyset) - E[E(Y|U, T = 0, F_T = \emptyset)|T = 1, F_T = \emptyset]. \qquad (6.7)$$

This is because $E(Y|U, F_T = 0) \equiv E(Y|U, T = 0, F_T = 0)$ can be rewritten as $E(Y|U, T = 0, F_T = \emptyset)$ by (6.3). Further $E[E(Y|U, F_T = 1)|T = 1, F_T = \emptyset]$ can be rewritten as $E(Y|T = 1, F_T = \emptyset)$ by (6.3).

Briefly, this relates to the ETT given in (6.1) in potential response terms as follows. The first part of (6.7) can be written as $E(Y_1|T = 1, F_T = \emptyset)$ as $T = 1$ implies that the potential response is Y_1 by definition (see section 3.2). The inner expectation of the second part is just Y_0 as $T = 0$, this the whole of the second part becomes $E(Y_0|T = 1, F_T = \emptyset)$. So (6.7) is

$$\begin{aligned} E(Y_1|T = 1, F_T = \emptyset) \ &- \ E(Y_0|T = 1, F_T = \emptyset) \\ \equiv E(Y_1|T = 1) \ &- \ E(Y_0|T = 1) \end{aligned} \qquad (6.8)$$

if we drop the intervention node notation. Now (6.1) is conditioned on an extra variable, X, and it would suffice to add this conditioning in (6.8) for the two to be the same (assuming that the conditional independences relating X to the other variables allow this).

The role of U

We have seen that the definition of the ETT depends on the choice of U. This has various consequences. Firstly, if U is observable, then the conditional independences (6.2) and (6.3) identify it as a sufficient covariate (Lauritzen 2001) for the effect of T on Y. However, as U is generally unknown, it acts as a confounder of the effect.

Further, if U is observable, then we can evaluate the average treatment effect from observational data. This is simply done by summing out over U.

$$
\begin{aligned}
&E(Y|F_T = 1) - E(Y|F_T = 0) \\
&\quad = \sum_u E(Y|U = u, F_T = 1)p(U = u|F_T = 1) - E(Y|U = u, F_T = 0)p(U = u|F_T = \\
&\quad = \sum_u \{E(Y|U = u, T = 1, F_T = \emptyset) - E(Y|U = u, T = 0, F_T = \emptyset)\}p(U = u|F_T =
\end{aligned}
$$

As $U \perp\!\!\!\perp F_T$ and $T = F_T$ when $F_T \neq \emptyset$, and $Y \perp\!\!\!\perp F_T|T, U$. All these quantities are observable and hence we would be able to identify and estimate the average treatment effect. We could also proceed to estimate the ETT, however, there would be no point to this exercise as we seek to estimate the ETT precisely because we are unable to estimate the ACE.

The choice of U appears to be somewhat arbitrary. If we are to try and evaluate the ETT, we must be certain that it is a well defined quantity, and that it does not depend on the choice of U.

To see this problem recall Doctor A in hospital One. He treated his patients according to some unknown criteria U. Consider now another doctor, Doctor B. Like his colleague Doctor A, he is given the drug to administer to the same patients. He bases his decisions on whether to treat a patient or not on his own set of criteria V, unknown to anyone else. Consider yet another doctor, Doctor C practising at the same hospital. He chooses to randomise treatment. In this case, his selection criterion corresponds to the empty set (still a possible choice) and the ETT is equivalent to the ACE.

The problem is whether the *choice of U changes the ETT given that the observable distributions for different choices of U are the same.* To clarify: Typically, different hospitals with different regimes will have different Us and thus different ETTs. This does not mean that U is not well-defined. The question we are asking is if in the *same (or exchangeable)* hospital, under the *same observable regime* and thus with the *same observable distributions* different choices of U will lead to different ETTs.

In order for the ETT to be well-defined, it cannot depend on the choice of U. It is constrained by the observable distributions which must be the same for different choices of U, thus we must determine whether, given the observed distributions are the same, the ETT remains the same irrespective of different choices of U.

If the ETT is the same for different Us then, we can choose U to be any variable that obeys conditional independences (6.2) and (6.3) and estimate the ETT.

6.2.1 Constraints

Let U and V be two different variables such that we can assume that conditional independences (6.2) and (6.3) hold for both. Generally, U and V are unobserved. In order for the ETT to be well defined, the joint distributions of the observed variables $(Y, T|F_T)$, which are the observable distributions, must be the same for both U and V for all regimes of F_T. This imposes certain constraints on these distributions. Note that the ETT depends explicitly on U but the observable distributions do not.

Consider the conditional independence structures described by the DAG in Figure 6.1 in section 6.2. If we imagine the same DAG with V replacing U, then denote the joint distributions for the four variables in the two situations (one with U and the other identical but with V replacing U) are

$$f_1(Y, T, U|F_T)$$

and

$$f_2(Y, T, V|F_T)$$

respectively.

The following must hold for the observable distribution.

C1 $f_1(Y, T|F_T = \emptyset) \equiv f_2(Y, T|F_T = \emptyset)$

C2 $f_1(Y,T|F_T=1) \equiv f_2(Y,T|F_T=1)$

C3 $f_1(Y,T|F_T=0) \equiv f_2(Y,T|F_T=0)$

Consider the scenario with U under the idle regime $F_T = \emptyset$. The joint distribution of $(Y,T|F_T=\emptyset)$ is as follows.

$$f_1(Y,T|F_T=\emptyset) = \sum_{u\in\mathcal{U}} f_1(U=u)f_1(T=t|U=u,F_T=\emptyset)f_1(Y|T=t,U=u) \quad (6.9)$$

For U with $F_T = 1$ say (as we assume that T can take on values 0 and 1) the joint distribution of Y and T reduces to just the distribution of Y as T is no longer uncertain.

$$f_1(Y|F_T=1) = \sum_{u\in\mathcal{U}} f_1(U=u)f_1(Y|T=1,U=u). \quad (6.10)$$

For U with $F_T = 0$ a similar expression is obtained.

$$f_1(Y|F_T=0) = \sum_{u\in\mathcal{U}} f_1(U=u)f_1(Y|T=0,U=u). \quad (6.11)$$

Similar expressions for the second scenario are obtained by replacing subscripts 1 with 2 and U with V.

It follows from the constraints listed above that we must require:

$$\begin{aligned} &\sum_{u\in\mathcal{U}} f_1(U=u)f_1(T=t|U=u,F_T=\emptyset)f_1(Y|T=t,U=u) \\ \equiv &\sum_{v\in\mathcal{V}} f_2(V=v)f_2(T=t|V=v,F_T=\emptyset)f_2(Y|T=t,V=v) \end{aligned} \quad (6.12)$$

and

$$f_1(Y|F_T=1) = f_2(Y|F_T=1) \quad \text{and} \quad (6.13)$$

$$f_1(Y|F_T=0) = f_2(Y|F_T=0). \quad (6.14)$$

Given these constraints, what can we say about the question at hand?

6.2.2 Initial Conjecture

In order to determine whether the ETT is well defined for arbitrary choices of U, we consider V, another variable obeying the same set of conditional independences w.r.t T, F_T and Y, and try and prove that under the constraints imposed by the observed distributions discussed in the previous section,

$$E[ACE(U)|T = 1, F_T = \emptyset] \equiv E[ACE(V)|T = 1, F_T = \emptyset]. \tag{6.15}$$

Again we are interested in the results of the same hospitals with the same regime and the same observable distributions. We would not expect the ETT to be the same from one hospital to another.

In the next section we show that when U is a pair of potential responses and V is another pair of potential responses, then the constraints above force the ETT to be the same. In section 6.4 we show that any U can be replaced by a pair of potential responses. It follows therefore that the constraints imposed by the equivalence of the observed distributions lead to the same ETT and hence, the ETT is well defined.

6.3 Potential responses Setup

Potential responses are used here as a mathematical tool to prove that the ETT is well defined.

Consider variables F_T, T, Y_1, Y_0, Y and V, where Y is a deterministic

function of T,Y_1 and Y_0 as follows:

$$Y \equiv \begin{cases} Y_1 & \text{if } T = 1 \\ Y_0 & \text{if } T = 0 \end{cases} \tag{6.16}$$

First we assume that

$$(Y_1, Y_0) \perp\!\!\!\perp F_T. \tag{6.17}$$

Further define $U = (Y_1, Y_0)$ and note that

$$U \perp\!\!\!\perp F_T \qquad \text{by (6.17), and} \tag{6.18}$$

$$Y \perp\!\!\!\perp F_T|(U,T) \qquad \text{by (6.16).} \tag{6.19}$$

Y_1 and Y_0 can be considered potential responses as defined by Rubin in 3.2.

Conditional independence (6.18) tells us that the existence and the joint distribution of Y_1 and Y_0 does not depend on whether an intervention takes place.

Conditional independence (6.19) says that the distribution of Y does not depend on whether an intervention took place or T arose naturally given U and the realized value of T.

Let V denote another variable that respects the same set of conditional independences w.r.t T, F_T and Y as U, (6.18) and (6.19).

6.3.1 Conjecture in terms of Potential responses

The initial conjecture given in (6.15) is equivalent to

$$E(Y|T = 1, F_T = \emptyset) - E[E(Y|U, F_T = 0)|T = 1, F_T = \emptyset] \equiv$$
$$E(Y|T = 1, F_T = \emptyset) - E[E(Y|V, F_T = 0)|T = 1, F_T = \emptyset] \tag{6.20}$$

which in turn is equivalent to

$$E(Y|T=1,F_T=\emptyset) - E[E(Y|U,T=0,F_T=\emptyset)|T=1,F_T=\emptyset]. \quad (6.21)$$

In terms of potential responses, this is

$$E(Y_1|T=1,F_T=\emptyset) - E(Y_0|T=1,F_T=\emptyset). \quad (6.22)$$

6.3.2 Constraints in potential response setup

Consider the following; Let $U = (Y_0, Y_1)$, a pair of potential responses, as defined in the section 4, *with some (as yet undefined) and unobservable* correlation structure. For F_T idle and $T = 1$, (6.9) becomes;

$$\begin{aligned} f_1(y,t=1|F_T=\emptyset) &= \sum_{y_0,y_1} f_1(y_0,y_1) f_1(t=1|y_0,y_1,\emptyset)\delta(y_1=y) \\ &= \sum_{y_0} f_1(y_1=y,y_0) f_1(t=1|y_1=y,y_0,\emptyset) \\ &= \sum_{y_0} f_1(y_0|y_1=y) f_1(t=1|y_1=y,y_0,\emptyset) f_1(y_1=y) \\ &= f_1(y_1=y) f_1(t=1|y_1=y,\emptyset). \end{aligned} \quad (6.23)$$

For $F_T = 1$ the above becomes

$$f_1(y_1=y). \quad (6.24)$$

Similary;

$$f_1(y,t=0|F_T=\emptyset) = f_1(y_0=y) f_1(t=0|y_0=y,\emptyset). \quad (6.25)$$

For $F_T = 0$ this becomes;

$$f_1(y_0=y). \quad (6.26)$$

Now, let $V = (W_0, W_1)$ be another set of potential responses obeying the set of conditional independences (6.17), (6.18), (6.19) and such that $U \neq V$, that is W_0 and W_1 have a different correlation structure than Y_1 and Y_0. It is possible to derive expressions like (6.23), (6.24), (6.25) and (6.26) by replacing f_1 with f_2 and y_0 with w_0 and y_1 with w_1.

Now the constraints are on the distributions that are observable in principle, this both idle and interventional regimes of F_T. Thus the constraints in the potential response setup are:

P1 $f_1(T|F_T = \emptyset) = f_2(T|F_T = \emptyset)$,

P2 $f_1(Y_0|T = 0, F_T = \emptyset) \equiv f_2(W_0|T = 0, F_T = \emptyset)$,

P3 $f_1(Y_1|T = 1, F_T = \emptyset) \equiv f_2(W_1|T = 1, F_T = \emptyset)$,

6.3.3 Proof of conjecture in terms of potential responses

Consider (6.22), this expression depends on the distribution of Y_0 given treatment is $T = 1$, and on the marginal distribution of Y_1

Consider the first part of (6.22). Now, $p(Y_1|T = 1, F_T = \emptyset)$ is an observable distribution and must be the same for W_1 by constraint P3. It remains to be seen if the constraints ensure that $f_1(Y_0|T = 1, F_T = \emptyset) = f_2(W_0|T = 1, F_T = \emptyset)$.

Denote by $p_t = p(T = t|F_T = \emptyset)$, the probability of $T = t$ given $F_T = \emptyset$. Assume that $p_t \neq 0$ for $t \in \{0, 1\}$. The probability p_t is the same for f_1 and

f_2 by constraint P1 above. Consider the following;

$$\begin{aligned} f_1(Y = y|F_T = 0) &= f_1(Y_0 = y|F_T = 0) \\ &= f_1(Y_0 = y|F_T = \emptyset) \qquad \text{as } (Y_0, Y_1) \perp\!\!\!\perp F_T \\ &= \sum_{t \in T} f_1(Y_0 = y|F_T = \emptyset, T = t) p_t \\ &= f_1(Y_0 = y|F_T = \emptyset, T = 1) p_1 + f_1(Y_0 = y|F_T = \emptyset, T = 0) p_0 \\ &= f_1(Y_0 = y|F_T = \emptyset, T = 1) p_1 + f_1(Y = y|F_T = \emptyset, T = 0) p_0, \end{aligned}$$

as when $T = 0$, $Y = y$. So

$$\begin{aligned} \overbrace{f_1(Y = y|F_T = 0)}^{(a)} &= \underbrace{f_1(Y_0 = y|F_T = \emptyset, T = 1)}_{(b)} \overbrace{p_1}^{(c)} \\ &+ \underbrace{f_1(Y = y|F_T = \emptyset, T = 0)}_{(d)} \overbrace{p_0}^{(e)} \end{aligned}$$

The above argument is valid for f_2 and W_0, W_1. Now (a), (c), (d) and (e) are observable quantities whose distributions given F_T must be the same for both U and V by constraints P1-P3. In particular, (a) is observable from experiment where we set $F_T = 0$, and (c),(d) and (e) from observation. Thus these are the same for the two sets of potential responses.

We can therefore solve for (b), and it also must be the same for both U and V. So we have that the constraints guarantee that $f_1(Y_0|T = 1) = f_2(W_0|T = 1)$. Thus we have shown that for any pair of potential responses, the constraints on the observable distributions guarantee that the ETT is the same.

6.4 Proof of conjecture for arbitrary U

To complete the proof that the ETT is well defined for arbitrary choice of U we show first that Y can be transformed into a variable that has the same probability distribution as Y and that is a deterministic function of U,T and a uniform random variable E that is independent of U and T given F_T. We do this using the *conditional probability transform* (Rosenblatt 1952). This function is then further shown to be an expression for a pair of potential responses $W = (Y_1, Y_0)$, such that $(U, E) \perp\!\!\!\perp Y | (W, T)$, $W \perp\!\!\!\perp F_T$ and further, $T \perp\!\!\!\perp W | U$. As the observable probability distributions given these transformations remain the same, the ETT remains the same.

The process is easily described as a sequence of graphs as shown in figure 6.2.

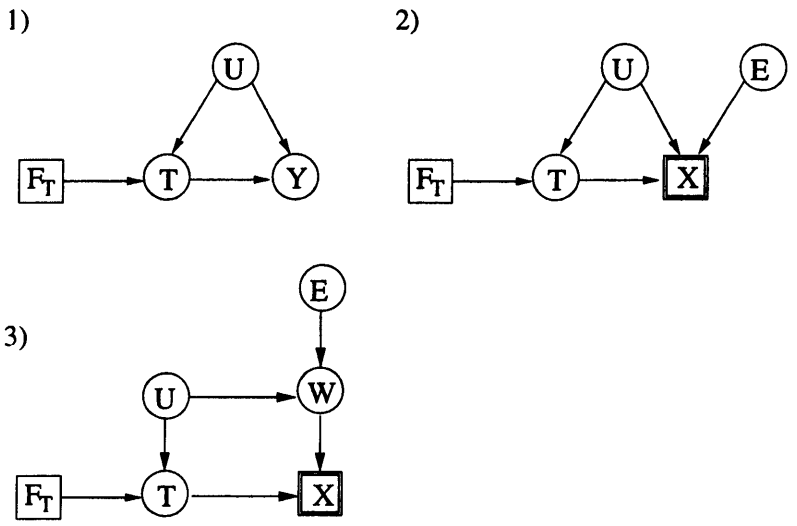

Figure 6.2: The transformation of Y into X. X is in a double box because it is a deterministic function of its parents.

Denote by

$$G_{tu}(y) = p(Y \leq y | T = t, U = u, F_T = \emptyset), \tag{6.27}$$

the cumulative distribution of Y given T and U. Assume that G_{TU} is well behaved, that is, Y is a continuous one dimensional variable. We must further require that G be non-zero inside its support. G can be extended to include discrete distributions as well. Then define

$$X := G_{TU}^{-1}(E) \tag{6.28}$$

where $E \perp\!\!\!\perp (T, U, F_T)$ and $E \sim \mathbf{U}[0, 1]$. Then for $t \in \{0, 1\}$ and $u \in \mathcal{U}$,

$$\begin{aligned} p(X \leq x | T = t, U = u, F_T = \emptyset) &= p(G_{tu}^{-1}(E) \leq x | T = t, U = u, F_T = \emptyset) \\ &= p(G_{tu}^{-1}(E) \leq x) \qquad \text{as } E \perp\!\!\!\perp (T, U, F_T) \\ &= p(E \leq G_{tu}(x)) = G_{tu}(x) \\ &= p(Y \leq x | T = t, U = u, F_T = \emptyset). \end{aligned}$$

Thus Y and X have the same conditional probability distributions, and X is a function of T, U and E. This implies further, that the joint distribution of X,T and U given $F_T = \emptyset$ is the same as the joint distribution of Y, T and U given $F_T = \emptyset$, which is what we require. As Y and X are both independent of F_T conditional on T and U, the conditional distributions of Y and X given $F_T \neq \emptyset$ are also equal to each other, and so the transformation is valid for all regimes of F_T. This step of the proof corresponds to DAG 2) in figure 6.2. Now write (6.28) as

$$X = g(T, U, E).$$

Further define $W = (Y_1, Y_0)$ where

$$Y_1 := g(1, U, E) \tag{6.29}$$

$$Y_0 := g(0, U, E). \tag{6.30}$$

Now Y is distributed like X which is such that

$$X = Y_1 \qquad \text{if } T = 1$$

$$X = Y_0 \qquad \text{if } T = 0$$

Hence W is a pair of potential responses as either one or the other will be observed given a treatment is administered.

We have shown that Y can be expressed in terms of a pair of potential responses that have the same conditional distributions as Y given T and U. The last part of the proof is to show that

1. $X \perp\!\!\!\perp (U, E) | (W, T)$,

2. $T \perp\!\!\!\perp (E, W) | U$ and

3. $W \perp\!\!\!\perp F_T$,

as this would imply that we can replace U and E with T and W in order to make inference about X and thus Y. The first conditional independence follows from the fact that X can now be expressed as a function of T and W: $X = g'(T, W)$ such that $X \equiv Y_t$ for $T = t$. Hence it is independent of U and E given T and W. For 2. we see that as $E \perp\!\!\!\perp T, U$, we have that $E \perp\!\!\!\perp T | U$. Further, as W is a function of E and U it follows that $T \perp\!\!\!\perp W | U$ by Dawid

(1979) Lemma 4.2. The last conditional independence follows from the fact that W depends only on U and E, both of which are marginally independent of F_T and hence W is also marginally independent of F_T. The last two steps correspond to DAG 3) in figure 6.2 and can be read off that graph using the moralisation criteria.

Hence we have shown that an arbitrary U can be expressed in terms of a pair of potential responses such that the probability distribution of Y given T and U remains the same throughout. Thus if we consider (6.7) which is one way of expressing the ETT in non-counterfactual terms;

$$E(Y|T=1,F_T=\emptyset) - E[E(Y|U,T=0,F_T=\emptyset)|T=1,F_T=\emptyset].$$

and (6.22)

$$E(Y_1|F_T=\emptyset) - E(Y_0|T=1,F_T=\emptyset).$$

we see that the first parts of the two equations is the same as $Y \sim Y_1$ when $T=1$ and so $E(Y|T=1,F_T=\emptyset) \equiv E(Y_1|F_T=\emptyset)$. Now $p(Y|U,T=0,F_T=\emptyset) \equiv p(X|U,T=0,F_T=\emptyset)$ by construction of X in (6.28), and $p(X|U,T=0,F_T=\emptyset) \equiv p(Y_0)$ by construction of Y_0 in (6.30). Hence the expectations also remain the same and there is equivalence between (6.7) and (6.22).

By the proof in section 6.3.3 we know that any two pairs of potential responses lead to the same value for the ETT given the constraints imposed by the observed distributions. Hence, as any U can be shown to lead to the same ETT as a pair of potential responses it follows that the ETT is

well defined for arbitrary choice of U given the constraints imposed by the observed distributions.

6.5 New Story

The first part of this chapter showed that given the constraints on the observable distributions, choice of U is not important and the ETT is well defined. In this part, we discuss the problem of identifiability of the ETT from both experimental and observational data.

Look again at the original story: Doctor A in hospital One is given a new drug treatment to administer to his patients. He observes some patient characteristics and makes a decision as to whether or not to administer the drug or not. This time however, the doctor's decision can be overruled and a different treatment to the one the doctor recommended can be administered. For example if the hospital runs a clinical trial. The doctor's decision D, and the treatment administered T as well as the responses Y are all recorded.

The statistician analysing the data is still ignorant of the doctor's decision criteria U. However, she knows what the doctor recommended, that is, D. We regard D as a chance variable arising naturally from its relationship with U. F_T is an intervention variable, and is equal to $\emptyset$ if the doctor administers the drug as he sees fit. In this case $T \equiv D$. If however, the hospital is running clinical trial, then $F_T = t$, $t \in \{0, 1\}$, and $T \equiv F_T$, regardless of the value of D. We see that the treatment that is actually administered is a deterministic

function of D and F_T as follows;

$$T \equiv \begin{cases} T = d & \text{iff } F_T = \emptyset \text{ and } D = d \\ T = t & \text{iff } F_T = t. \end{cases}$$

The variables above are related as shown in the DAG 6.3. In particular,

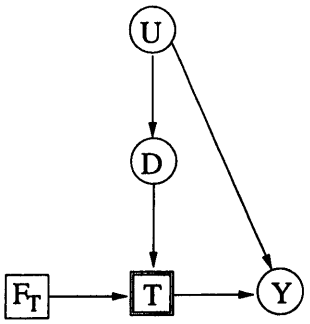

Figure 6.3: The story in a DAG.T is in a double square as it is in fact a deterministic function of D and F_T.

the DAG exhibits the following set of conditional independence constraints;

$$\begin{aligned} (U, D) &\perp\!\!\!\perp F_T \\ Y &\perp\!\!\!\perp (D, F_T)|(U, T) \\ T &\perp\!\!\!\perp U|(F_T, D). \end{aligned}$$

The last conditional independence is trivial, as T is a deterministic function of F_T and D. From these we can derive (using moralisation criterion 2.4) that

$$D \perp\!\!\!\perp F_T \tag{6.31}$$

$$Y \perp\!\!\!\perp F_T|(T, D), \tag{6.32}$$

In other words, we can replace U with D in the ETT as knowing U gives us no extra information about the situation if we already know D. Hence (7) can be re expressed in terms of an observable value.

Going back to equation (6.4) and replacing U with D, what does the ETT turn into?

$$\begin{aligned} & E[E(Y|D, F_T = 1)|T = 1, F_T = \emptyset] \quad - \quad E[E(Y|D, F_T = 0)|T = 1, F_T = \emptyset] \\ = & E[E(Y|D, F_T = 1)|T = 1, F_T = \emptyset] \quad - \quad E[E(Y|D, F_T = 0)|D = 1, F_T = \emptyset] \\ & = E(Y|D = 1, F_T = 1) - E(Y|D = 1, F_T = 0). \end{aligned} \tag{6.33}$$

This is because $T = 1$ and $F_T = \emptyset$ is equivalent to $D = 1$ and F_T. As well as being a simpler version of (6.4), it also has a very simple interpretation. It is the difference of the expectation of the response between those that were treated and those that were not treated given that the doctor would have liked to treat them all. The reason for using this version is to simplify the problems of identification from observational data as we will see in section 6.6.

Econometric Choice Model

The classic econometrics choice model (Heckman and Navarro-Lozano 2004) is analogous to the above story. In this model, a group of people decide whether they want to join a particular scheme. However, only a subset of them actually participate, and the reasons they made the decision to participate are unknown. Again we seek to estimate the ETT.

In the doctor example the population is that of the patients of the hospital, and is hence a closed and easy to monitor population. In the choice model setting, the population of interest is in principle, the whole population,

which is difficult to monitor. The solution to this problem is to consider a subset of people in the general population that has certain characteristics of interest. The probability that someone in this subpopulation would choose to participate, the *propensity score* (Rosenbaum and Rubin 1983), is found by means of surveys, and is generally considered known.

Where in the doctor example everything is conditional on the hospital (although this is not made explicit), in the choice model, we condition on the covariates that delimit the population of interest, say X. Then, the effect of treatment on the treated is given by

$$E(Y|D=1,T=1,X) - E(Y|D=1,T=0,X). \tag{6.34}$$

Consider the following example to clarify the interpretation of (6.34) above: Legal immigrants in the north of Italy can sign up to participate in a programme that trains them to set up a business. A number of them are randomly selected to participate in the training scheme. After the end of the course those who participated are monitored for an additional year and their success in setting up or participating in a business venture recorded. Y is the response, the participant's progress in the business, D is the variable that tells us whether they signed up to participate. D is 1 if the individual wanted to participate and 0 if they did not. T is 1 if the individual participated and otherwise. Now (6.34) is equal to $E(Y|D=1,F_T=1,X) - E(Y|D=1,F_T=0,X)$ provided $Y \perp\!\!\!\perp F_T|D,T,X$ which is assumed to be the case. Thus we find that we are again in the situation we have been addressing so far.

6.6 Identification

We now look at the problem of whether the ETT is identifiable from experimental data where $F_T \neq \emptyset$ or from observational data, where $F_T = \emptyset$, or even a combination of the two. We return to the situation where we cannot observe D. It is easy to see that $E(Y|D = 1, F_T = 1)$ is identifiable from the observational data, as it is equivalent to $E(Y|D = 1, T = 1, F_T = \emptyset)$ by the independence constraint (6.32) and these are exactly the conditions for recording the data. $E(Y|D = 1, F_T = 0)$ however poses more of a problem. Consider $p(Y|F_T = 1)$, the response given $F_T = 1$, under experimental conditions, where we just consider the response of those who were treated, without considering the doctor's recommendation.

$$
\begin{aligned}
p(Y|F_T = 1) &= \sum_{d \in \mathbf{D}} p(Y|D = d, F_T = 1) p(D = d|F_T = 1) \\
&= p(Y|D = 0, F_T = 1) p(D = 0|F_T = \emptyset) && (6.35) \\
&\quad + \underbrace{p(Y|D = 1, F_T = 1)}_{(a)} \underbrace{p(D = 1|F_T = \emptyset)}_{(b)} && (6.36) \\
&= p(Y|D = 0, F_T = 1) p(T = 0|F_T = \emptyset) && (6.37) \\
&\quad + \underbrace{p(Y|D = 1, T = 1, F_T = \emptyset)}_{(c)} \underbrace{p(T = 1|F_T = \emptyset)}_{(d)} && (6.38)
\end{aligned}
$$

(6.35) to (6.37) follows from the fact that when $F_T = \emptyset$, $D \equiv T$. (a) in (6.36) becomes (c) in (6.38) by conditional independence (6.32). (b) in (6.36) and (d) (6.38) are equivalent because when $F_T = \emptyset$, $D \equiv T$.

Looking at the set of equations above, we see that whereas $p(Y|F_T = 1)$ can only be estimated directly by an experiment where all patients are admin-

istered treatment $T = 1$, the remaining probabilities except the problematic $p(Y|D = 0, F_T = 1)$, can in principle be estimated by letting the doctor administer the treatment as he chooses. In this case, $D \equiv T$ and $F_T = \emptyset$ and we are in the observational regime.

In order to identify $p(Y|D = 0, F_T = 1)$, we would therefore have to have *two* exchangeable groups of patients, one to whom administer treatment $T = 1$ and one to observe a doctor's behaviour.

So far, we have looked at Doctor A in hospital One. Let us assume that there is another hospital, hospital Two, whose patients are exchangeable with those in hospital One, where Doctor A is making the same decisions D as he did in hospital One, and administering treatments accordingly.

Now $p(Y|F_T = 1)$ can be found by running a clinical trial and randomising patients in hospital One. $p(T = 0|F_T = \emptyset)$ can be found from an observational study, where the doctor B is left to his own devices in hospital Two, as $p(T = 0|F_T = \emptyset) \equiv p(D = 0|F_T = \emptyset)$. Similarly $p(T = 1|F_T = \emptyset)$ and $p(Y|D = 1, T = 1, F_T = \emptyset)$ can be found from leaving the doctor to administer the drug.

It would appear that if we can run an experiment there would be no point in evaluating the ETT, as we could evaluate the ACE directly, however identification of $p(Y|D = 0, F_T = 1)$ requires only $p(Y|F_T = 1)$ *not* $p(Y|F_T = 0)$, and hence, a complete randomised clinical trial is not necessary. It is only necessary that hospital One administer treatment $T = 1$ to all or a random sample of its patients indiscriminately. Thus it is possible to identify the ETT

if there is no control group. These are very specific circumstances and are not likely to occur in reality. It is usually the case that there is no experimental data and thus additional assumptions must be made.

In the econometrics choice model, we do not have experimental data as no one can be forced to do anything against their will, training programmes have a limited number of places etc. So the problem becomes how to identify the ETT from purely observational data. In order to do this, different techniques, involving different initial assumptions have been developed. These are in particular the assumptions of *matching*, see Rosenbaum and Rubin (1983) and (1985), *control functions*, Heckman and Robb (1985) and finally *instrumental variables*, Angrist et al. (1996). A review of these methods and extensive bibliography are given in Heckman and Navarro-Lozano (2004).

I will discuss matching and control functions and their non-counterfactual counterparts.

6.6.1 Identification assumptions in the potential response framework

This section discusses two of the assumptions made for identifying the ETT from observational data in potential response setups, Matching and Control functions (Heckman and Navarro-Lozano 2004). In order to do this their initial setup, which is slightly more complex than the one developed in the non-counterfactual framework so far must be described.

Heckman's model of choice has the following storyline: a sub-population

determined by some observable characteristics can decide whether they *would like* to participate in a scheme, for a number of reasons, they may not necessarily be *able* to participate in the scheme. Hence the attention is restricted by necessity, to the sub-population of those who did participate, i.e. those who were treated. The quantity of interest is the effect of treatment on the treated. The model has the following elements;

List 6.6.1

I1 T, the decision variable, not to be confused with D the decision variable used in the non-counterfactual setup.

I2 $Y = (Y_0, Y_1)$, the potential outcome variables. $Y = Y_t$ iff $T = t$, $t = 0, 1$.

I3 Z, a set of observed variables known to (at least partially) influence the choice.

I4 X, a set of observed covariates such as age, sex and some demographics.

I5 U_0 and U_1, unobserved variables that influence Y_0 and Y_1 respectively through the functions $Y_t = \mu_t(X, Z, U_t)$, $t = 0, 1$.

I6 U_V, unobserved factors affecting choice, these are related to a utility function $V = \mu_V(X, Z, U_V)$. Further, each individual will choose $T = 1$ (given X and Z) if and only if they believe that its utility V is greater or equal to 0.

I7 The probability $p(T = 1|X, Z)$, known as the *propensity score*, the probability that an individual in a particular subpopulation determined

by the values of X and Z chooses to participate in the scheme. The propensity score is generally considered known (through surveys and census information) and is assumed to depend entirely on X and Z. It will be denoted from here on as $ps(X, Z)$.

The effect of treatment on the treated is given by

$$ET = E(Y_1 - Y_0|T = 1, X) \equiv E(Y_1|T = 1, X) - E(Y_0|T = 1, X). \quad (6.39)$$

The following quantities are assumed observable; $E(Y_1|T = 1, X, Z) = E(Y|T = 1, X, Z)$ and $E(Y_0|T = 0, X, Z) = E(Y|T = 0, X, Z)$. Also the expectations of Y_t given $T = t$, that is $p(T = 1|X, Z)$, $p(T = 0, X, Z)$ the propensity scores are assumed known. Finally, $p(Z|X, T)$, the probabilities of T given the observed covariates, and the distribution of $Z|X, T$ are also assumed to be observable. As with the non-counterfactual setup the problem is identifying $E(Y_0|T = 1, X)$.

Matching

In matching the assumption is made that $(Y_1, Y_0) \perp\!\!\!\perp T|(X, Z)$. This then means that

$$E(Y_0|T = 1, X, Z) = E(Y_0|T = 0, X, Z) = E(Y_0|X, Z). \quad (6.40)$$

In fact, for the effect of treatment on the treatment, the assumption can be reduced to $Y_0 \perp\!\!\!\perp T|(X, Z)$. This is in fact analogous to randomising the treatment given (X,Z). If we consider Y_0, we know this means that the treatment

that has been administered is 0. In non-counterfactual terms, this can be seen as saying that $F_T = 0$. The conditional independence (6.40) can be reinterpreted as $Y \perp\!\!\!\perp F_T | X, Z, T$. This assumption equates the effect of treatment on the treated to the average treatment effect, and is not often appropriate.

Control Functions

The control function approach relies on the assumption of *separability* of each potential response into two parts that can be treated independently of one another. Usually this assumption is strengthened to *additive separability*[1] which means that the functions in items (I5) and (I6) in list 6.6.1 above can be expressed as follows:

$$Y_t = \mu_t(X, Z) + U_t, \tag{6.41}$$

$$V = \mu_V(X, Z) + U_V, \tag{6.42}$$

where we assume that $(U_1, U_0, U_V) \perp\!\!\!\perp (X, Z)$. This guarantees that the two parts the right hand sides of (6.41) and (6.42) are composed of can be treated separately. From (6.41) it follows that

$$E(Y_t | T = t, X, Z) = \mu_t(X) + E(U_t | X, Z, T = t). \tag{6.43}$$

So, for the case where $T = 1$.

$$E(Y_1 | T = 1, X, Z) = \mu_1(X) + \underbrace{E(U_1 | X, Z, T = 1)}_{(a)}. \tag{6.44}$$

[1] In Cameron and Heckman (1998) there are examples where it is possible to avoid the additive part of the assumption.

Look at *(a)* in (6.44).

$$\begin{aligned}(\text{a}) = E(U_1|X,Z,T=1) &= E(U_1|V \geq 0, X, Z) \\ &= E(U_1|U_V \geq -\mu_V(X,Z), X, Z), \quad (6.45)\end{aligned}$$

by 6.42. Now we know X and Z (as these are observed). Further, for simplicity we say that $w = -\mu_V(X,Z)$. Then (6.45) is

$$E(U_1|U_V \geq w) = f(w) \tag{6.46}$$

a function of w. Similarly, the propensity score itself is a function of w as

$$\begin{aligned}p(T=1|X,Z) &= p(V \geq 0) \\ &= p(U_V \geq -\mu_V(X,Z)) \\ &= p(U_V \geq w) = g(w) \quad (6.47)\end{aligned}$$

Assume that we know the probability distribution g of U_v, and it is smooth and continuous. See figure 6.4. We have from (6.47) above that $p(T =$

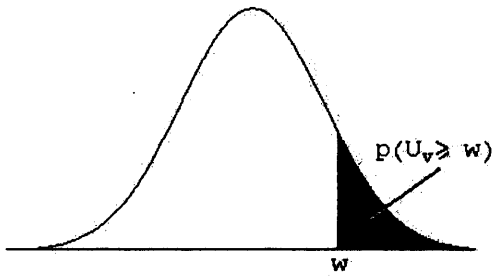

Figure 6.4: If we know the mass of the tail area probability $p(U_v \geq w)$ and the cumulative probability function we can infer w.

$1|X,Z) = g(w)$, so for a given X and Z we can invert g:

$$w = g^{-1}(p(T = 1|X, Z)). \tag{6.48}$$

Then if we know the joint probability of U_1 and U_V, we can calculate *(a)* in (6.44) above by plugging w found from (6.48) into (6.46), so

$$\begin{aligned} E(U_1|U_V \geq w) &= f(g^{-1}(p(T = 1|X, Z)). \\ &= K_1(p(T = 1|X, Z)) \\ &= K_1(ps(X, Z)). \end{aligned} \tag{6.49}$$

Where K_1 is just a function of the propensity score. As we do not know g or f in actuality situation, K_1 must be assigned a specific form and its components estimated as follows: $E(Y_t|T = t, X, Z)$ for both $t = \{0, 1\}$ can be estimated from the observed data for both settings of T and any available setting of X and Z. Further , the propensity score $ps(X, Z)$ is considered known or estimable for all available settings of X and Z. In order to estimate components of K_1 (as well as the μs the expectation of Y_t), $E(Y_t|T = t, X, Z)$ is regressed against the propensity score.

A similar argument is made for $T = 0$ and we get

$$E(U_0|X, Z, T = 0) = E(U_0|U_V < -\mu_V(X, Z)) = K_0(ps(X, Z)). \tag{6.50}$$

The μs estimated in this way can be identified up to a constant term. However, this constant term is not important to estimate the ET as it is assumed to cancel out when the expectation of Y_0 is subtracted from the expectation of Y_1.

Due to the assumption of separability, we can then take the μ we have estimated from treatment $T = 1$ and add it to K_0 to deduce

$$E(Y_0|T = 1, X, Z) = \mu_1(X) + K_0(ps(X, Z)).$$

The ET is more complex than the subtraction of the two expectations of Y, as Z needs to be summed out. The details are beyond the scope of this discussion. See Heckman and Vytlacil (2005) for such details.

6.7 Non-counterfactual assumptions for identification

It is not possible to identify $E(Y|D = 1, F_T = 0)$ from observational data, or indeed data involving interventions without control groups without imposing some additional assumptions. These are of a parametric form and are inspired by the counterfactual approaches.

6.7.1 Matching

In the non-counterfactual framework, matching is expressed as $Y \perp\!\!\!\perp D|(F_T, T)$. Hence ETT becomes

$$\begin{aligned}
&E(Y|D = 1, F_T = 1) - E(Y|D = 1, F_T = 0) \\
&\qquad = E(Y|D = 1, T = 1, F_T = 1) - E(Y|D = 1, T = 0, F_T = 0) \\
&\qquad = E(Y|T = 1, F_T = 1) - E(Y|T = 0, F_T = 0) \\
&\qquad = E(Y|F_T = 1) - E(Y|F_T = 0),
\end{aligned}$$

which is the average treatment effect. As in the counterfactual setup, this assumption is often not appropriate.

6.7.2 Control Functions

The idea follows that of control functions in the counterfactual framework, but is slightly different. First, we (re)introduce the unobserved variable U, and as we need to be able to condition on a set of observable covariates, we introduce a set of these denoted by Z. These are believed to influence choice, and if we adopt the utility argument put forward by Heckman, the decision to participate depends entirely on this set of variables. These variables are assumed to obey the following conditional independences:

Assumption 1

$$(U, Z) \perp\!\!\!\perp (F_T, T) | D, \tag{6.51}$$

and

Assumption 2

$$(U, Z) \perp\!\!\!\perp F_T \quad \text{and} \tag{6.52}$$

$$U \perp\!\!\!\perp Z. \tag{6.53}$$

A change in the conditional independence (6.3) involving Y also has to be made to account for the introduction of Z. It is changed to

$$Y \perp\!\!\!\perp F_T | (U, D, T, Z) \tag{6.54}$$

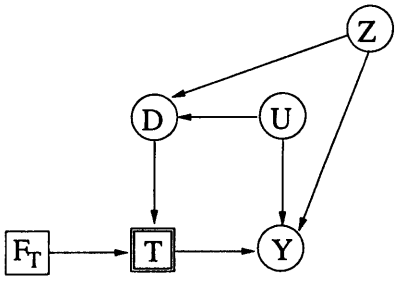

Figure 6.5: DAG represents the relationships between the variables. T is in a double box as it is a deterministic function of F_T and D.

These relationships are described in the DAG figure 6.5.

The first conditional independence says that both the unobserved variable U and the observed covariates are independent of the treatment assignment mechanism given the decision, i.e. the doctor's preferred treatment. The second conditional independence just states that the observed and unobserved covariates are marginally independent of the intervention variable.

The third conditional independence tells us that the observed and unobserved influences on the decision are independent of one another. This assumption may seem implausible, however, in the econometric context Z is generally considered to contain characteristics such as age and sex as well as demographic details, whereas U is considered to be either an error term or to contain personal detail that might influence each individual in his/her own way. This can be considered marginally independent of the demographic details. In the doctor story, U may be the doctor's personal feeling about who should be treated, and Z the patient characteristics, age, sex and medical history. The two are *not* independent given D however.

We further assume additive separability as follows;

Assumption 3

$$Y = \mu_T + U \tag{6.55}$$

This implies

$$\begin{aligned} E(Y|F_T = t, D = d, Z) &= \mu_t(Z) + E(U|F_T = t, D = d, Z) \\ &= \mu_t(Z) + E(U|D = d, Z) \end{aligned} \tag{6.56}$$

as $(U, Z) \perp\!\!\!\perp F_T|D$. $\mu_t(Z)$ is the part that depends on the value of F_T (and hence T) as well as the observed covariates. $E(U|D = d, Z)$ depends on D and Z through U.

Look at $E(U|D = d, Z)$ in more detail; if we make the further assumption that

Assumption 4

$E(U|D = d, Z)$ is a function of the probability that $D = d$ given the observed covariates and we denote $p(D = d|Z = z) = p_d(z)$ and say

$$E(U|D = d, Z = z)) = f(p_d(z)). \tag{6.57}$$

We can then determine $\mu_t(Z)$ by regressing Y (which we have) against $f(p_d(Z))$ for the cases where $F_T = D$ over the domain of Z.

We can justify assumption 4 using a utility based argument similar to Heckman's.

Assumption 5

An individual will choose $D = 1$ if and only if he thinks that it will have a higher utility than not making the choice.

We assume that we can estimate the probability of making the choice $D = 1$ from historical data or that it is known and depends only on the observed covariate. This is in fact the propensity score, called p_d here. Hence we assume that we know $f(p_d(z))$ for all available $z \in \mathcal{Z}$ where $\mathcal{Z}$ is the domain of Z.

Having estimated μ_1 and μ_0 we then need to sum out over Z in order to be able to estimate $E(Y|D = 1, F_T = 0)$. Once this has been done, we can estimate $E(Y|D = 1, F_T = 0)$ as follows;

$$\begin{aligned} E(Y|D = 1, F_T = 0) &= \mu_0 + E(U|F_T = 0, D = 1) \\ &= \mu_0 + E(U|D = 1) \\ &= \mu_0 + f_1(p_1). \end{aligned} \tag{6.58}$$

So we have identified the effect of treatment on the treated.

6.8 Conclusions

It is worth making some final comments on the discussion in this chapter. No assumption about the existence of potential responses were made or any constraints placed on their correlation structure in this chapter. They are used as a mathematical tool. As long as they have the same observable distributions, the unobservable joint distribution of the potential responses is irrelevant to the ETT. Of course the value of the ETT will depend heavily on

the parametric assumptions made in order to identify it, thus overriding the fact that the ETT as a mathematical expression is well-defined irrespective of the joint distribution of the potential responses.

Another interesting point is that when tackling this problem initially, we had considered only the idle regime of F_T. In so doing, it appeared immediately obvious that the ETT was not well defined, as a counterexample was produced that showed that different U's lead to different ETTs. It was only upon realising that the constraints had to include all distributions of the observable variables that were at least *in principle* observable, that is, including the interventional regimes, that it was possible to prove that the ETT was well-defined. This is interesting particularly in the context of potential responses and counterfactuals where the relationship between intervention and causality is not explicit in the notation or indeed in discussion of problems. Heckman himself makes the assumption that a potential response variable $Y_c(u)$ is the same irrespective of how c comes to be, by intervention or otherwise.

A final point is the relative ease with which counterfactual arguments can be turned into plain probabilistic arguments. This is a theme that has been widely explored in this thesis, and this chapter is yet another example of it.

Chapter 7

Conclusions and further work

Causal inference is a fascinating and controversial subject and can be approached from many different angles depending both on the background of the researcher and his or her perception of causality itself. As it is relatively new as a field of research in statistics, a unified approach to it has not yet been established. In this thesis we have proposed an approach based on decision theory which we believe to be optimal for predictive causal inference. This reflect the idea that causality can be understood best in terms of decisions and interventions. We have also avoided making untestable assumptions about unknowable quantities such as potential responses and tried where possible to express causal quantities in terms of probabilities without deterministic relationships.

We have explored and assessed the validity of assumptions made in competing causal models and argued that these are not useful or realistic and lead to errors both in how causal quantities are expressed and how the methods are applied.

In Chapter 4 we showed that misuse of the counterfactual/graphical set-up can lead to incorrect inference about the existence of causal relationships in the area of causal discovery from observational data, and clearly stated the implicitly made assumption that relationships derived from observational studies are the same as those derived from experiments.

In Chapter 5 we showed that the problem of direct and indirect effects could be expressed more simply in terms of the decision theoretic framework and introduced the fictional manipulation variable which codes random interventions and thus enables us to identify direct and indirect effects form observational data without making further assumptions.

In Chapter 6 we tackled the effect of treatment on the treated and showed that this effect is well-defined for any unobserved sufficient covariate. Further we showed that it is not possible to identify the ETT entirely from observational data without making further modelling assumptions.

We trust that the contents of this thesis has convinced the reader that this approach is superior to the competing causal frameworks for predictive causal inference.

Having set up the decision theoretic causal framework, further research is unlimited. One interesting avenue of research is that of blackbox interventions, applied in particular to policy interventions.

Another is to extend the concept of manipulation variables to problems other than that of direct and indirect effects. An example of such a problem, also termed the problem of direct and indirect effects, refers to the situation

where the treatment of one individual will have an effect directly on the individual itself as well as those who surround him or her.

An example of this type of problem is when a sample of a population are vaccinated. This vaccine has an effect on the vaccinated individual and on his community (family and collegues). This is a difficult problem as it involves looking at the effect of interventions on different levels, individuals on the one hand, and the community on another. It would therefore involve coding interventions on different levels too.

Remaining in the area of methodology, it would also be interesting to take a step back in the process and look at how to design experiments, or observational studies such as surveys to optimise the estimation of causal quantities.

Further, there is the possibility of modelling the difference between the data we observe and the data we need in order to make causal inference as a separate parameter (Greenland 2005).

There is also the problem when making causal inference using observational studies that there comes a point where additional information is no longer useful as the data is not experimental. It would therefore be interesting to explore the problem by including a *value of information* argument in the methodology.

Finally, it would be interesting to see how the methodology developed here works in a real context with data.

Appendix A

Markov Equivalence

Markov Equivalence in DAGs is characterised in the following way. Two DAGs G_1 and G_2 are Markov Equivalent if they have the same *skeleton* and the same *v-structures*. The skeleton of the DAG is the set of nodes and edges between them. Hence the three DAGs in Figure A.1 have the same skeleton. The v-structures of a DAG are the triples of variables, say (X, Y, Z) such that $X \rightarrow Y \leftarrow Z$. Hence in Figure A.1 DAGs 1) and 2) have the same v-structures, namely $B \rightarrow D \leftarrow C$ whereas DAG 3) has an additional v-structure $B \rightarrow A \leftarrow C$. In fact DAGs 1) and 2) are Markov equivalent and embody the same conditional independence relationships:

$$D \perp\!\!\!\perp A \,|\, (B, C)$$

$$B \perp\!\!\!\perp C \,|\, A$$

whereas DAG 3) embodies the following conditional independence relationships:

$$D \perp\!\!\!\perp A \,|\, (B, C)$$

$$B \perp\!\!\!\perp C$$

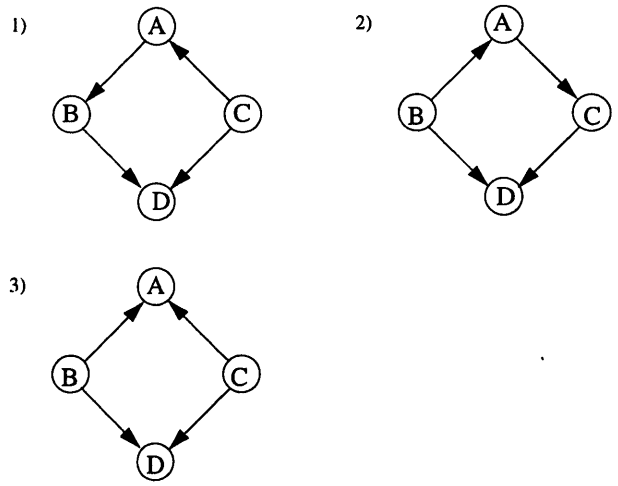

Figure A.1: Figures 1),2) and 3) have the same skeleton, but whereas 1) and 2) have the same v-structure and are therefore Markov Equivalent, 3) does not.

Appendix B

Simple Causal Discovery Algorithms

The IC Algorithm

The *Inductive Causation algorithm* can be found in (Pearl and Verma 1991) and (Pearl 2000) Chapter 2.5. A further algorithm called the *Inductive Causation algorithm with Latent variables* is aimed at discovering causal relation in cases where unobserved variables are believed to be parents of some of the observed variables. The IC algorithm works as follows:
We input $\hat{P}$ a stable distribution on a set of variables V.

1. For each pair of variables A and B in V, search for a set S_{AB} such that $A \perp\!\!\!\perp B | S_{AB}$ w.r.t the probability distribution $\hat{P}$. We insert an undirected edge between A and B if no set S_{AB} can be found.

2. For each pair of non-adjacent variables A and B with a common neighbour C, check if $C \in S_{AB}$.

   If it is then move on to step 3, otherwise add arrowheads pointing at

C.

3. In the partially directed graph resulting from the above two steps orient as many of the edges according to the following rules:

   (i) the orientation should not create new v-structures and

   (ii) the orientation should not create a directed cycle.

Example

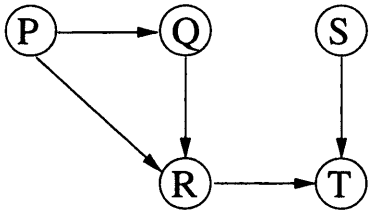

Figure B.1: Generating DAG

Imagine we have data generated by the DAG in Figure B.1. From the data we find the following set of conditional independences:

$P \perp\!\!\!\perp S$,

$P \perp\!\!\!\perp T | R$,

$Q \perp\!\!\!\perp S$,

$Q \perp\!\!\!\perp T | R$,

$R \perp\!\!\!\perp S$.

We then use the IC algorthim to find the generating graph. For Steps 2 and 3 below, I just show how the algorithm works for 2 sets of non-adjacent variables.

Step 1: put an undirected edge between each every pair of nodes that do not have a set that separates them. In this case the variables with an undirected edge are those that do not appear with a $\perp\!\!\!\perp$ between them in the above list. The result is Figure B.2 a), which has the same skeleton as the generating graph.

Step 2: the sets of non-adjacent variables are (P,T), (P,S), (Q,T), (Q,S) and finally (R,S). Consider (P,T): they have a common neighbour R. Is R in S_{PT}? It is, hence we continue to the next step. Consider (R,S), these have a common neighbour T. Is T in S_{RS}? No it is not, hence add arrows pointing at T from R and S. See Figure B.2 b).

Step 3: orient the graph as we choose providing we do not create cycles or v-structures. See Figure B.2 c) for the set of Markov Equivalent graphs generated by this process.

Six ME DAGs are generated by the last step. Only one corresponds to the data generating graph. We may be able to eliminate some of the DAGs if we have further information such as for example P precedes R. This would exclude 3 of the DAGs.

All the DAGs found using the IC algorithm share the v-structure. Hence, if we interpret the DAGs as causal, we can make inference about intervention on R and S and their effect on T as detailed in section 4.7. This can be extended to the general case. The sets of Markov equivalent DAGs generated from a set of conditional independences (with no additional constraints), will only enable causal inference on the v-structures, as these are the only directed

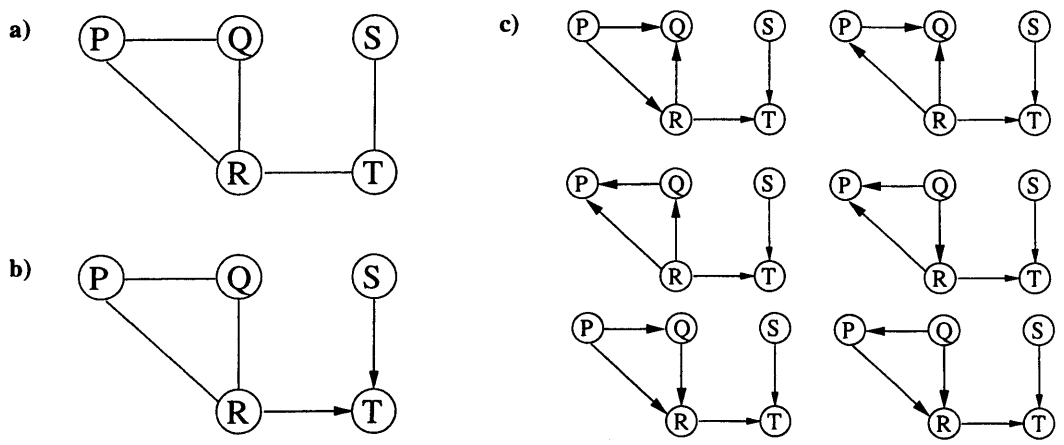

Figure B.2: The steps of the IC algorithm

elements that they have in common.

The IC algorithm would not be very useful in highly complex problems involving many variables that are related to each other. If we do not have sufficient v-structures, then this method would not result in very many relationships that can be used for inference. For example, if we had not had the conditional independence $R \perp\!\!\!\perp S$, then there would have been an extra edge between R and S and hence no v-structure that can be used for inference. The process allows DAGs to be excluded when there is prior information about temporal ordering or the variables or when variables are considered to be root variables and can therefore not have any parents (such as for example gender). The situation would be further complicated when we allow for latent variables.

Bayesian Causal Discovery

The Bayesian causal discovery method differs from the constraint based method in one main aspect. Whereas the constraint based methods take the conditional independences recovered from the data as true, the Bayesian method associates a degree of uncertainty to the constraints, by allowing different graphical models (which encode different conditional independences) to represent the observational data set. (Heckerman, Meek, and Cooper 1999) state that the Bayesian approach is superior to the constraint based method for three reasons. It is not subject to errors due to incorrect conditional independence constraints. Finer distinctions among model structures can be made , and finally, models can be combined to make better inference and take into account model uncertainty.

Outline of approach

1. Consider a set of variables $\mathbf{X} = \{X_1, \ldots, X_n\}$.

2. We have complete data $D = \{x_1, \ldots, x_n\}$, which is a random sample from some unknown probability distribution for $\mathbf{X}$.

3. We assume that this unknown probability distribution can be encoded by some causal model with structure $\mathbf{m}$.

4. Let $\mathbf{m}$ be a realization of the discrete $\mathbf{M}$, each $\mathbf{m}$ representing a possible true model.

5. Each $\mathbf{m}$ has associated uncertainty in the form of a probability distribution $p(\mathbf{m})$

6. Every possible model structure $\mathbf{m}$ has an associated set of parameters Θ_m, whose values θ_m have associated uncertainty $p(\theta_m|\mathbf{m})$.

7. Given D, compute the posterior probability for each $\mathbf{m}$ and θ_m, $p(\mathbf{m}|D)$ and $p(\theta_m|D,\mathbf{m})$ respectively using Bayes's rule.

8. Given a hypothesis h, determine probability that h is true given the data D by averaging over all possible models and their parameters

$$\begin{aligned} p(h|D) &= \sum_m p(\mathbf{m}|D)p(h|D,\mathbf{m}) \qquad \text{where} \\ p(h|D,\mathbf{m}) &= \int p(h|\theta_m,\mathbf{m})p(\theta_m|D,\mathbf{m})d\theta_m \end{aligned}$$

If we assume that the likelihood terms factorise into local groups, that is

$$p(\mathbf{x}|\theta_m,\mathbf{m}) = \prod_{i=1}^{n} p(x_i|pa_i,\theta_i,\mathbf{m})$$

and that each local likelihood is in the exponential family, and further that the parameters are mutually independent then the above computations can be done efficiently and in closed form.

An artificial example of how the algorithm works is given in (Heckerman, Meek, and Cooper 1999). It starts off by generating data from the model $X \rightarrow Z \leftarrow Y$ and then uses this data to test the hypothesis that "X causally influences Z". The model is just a graphical representation of the conditional independence $X \perp\!\!\!\perp Y$. The example goes on to find that the generating model

is indeed the most likely model for the observational data. This shows that the algorithm successfully finds its generating model, but does not justify causal interpretation of the models. This can only be done by making the assumption that there is an underlying causal structure and it represents both the experimental situations as well as the natural one. That is, the fundamental assumption.

I imagine that the limitations of the Bayesian approach are similar to those of the IC algorithms when it comes to large numbers of inter-related variables. However, these algorithms are not bound to the conditional independences as the IC algorithm is, and hence will probably be less prone to errors associated to finding the incorrect conditional independences.

In (Heckerman and Shachter 1995), a decision-theoretic approach to causality is described, which is covered in Chapter 3 section 4. It allows for variables that cannot be intervened upon to be decision variables. For example, although gender cannot physically be intervened upon, it could in theory at least, have been intervened upon and changed at conception. It is therefore possible to make inference with gender as a cause. This allows for a more liberal interpretation of discovered DAGs as encoding causal relationships.

Appendix C

Humans vs Animals

A new drug is found to be very effective in animals. The drug is very strong and has some unpleasant side-effects and the pharmaceutical company producing it would like to know how well its effect on animals predicts its effect on humans before it is tested on humans. Let T be the treatment, A be the response in animals and H the response in humans. See Figure C.1. The pharmaceutical company can test the low dosage on both humans and animals, however, the can only test the high dosage on animals. If the direct effect and indirect effect via animals can both be found for the low dosage, they expect the high dosage relationships to remain the same as they are considered stable.

Note that in this example, the relationship between A and H is purely associational, that is there is no reason or cause that links the two. Furthermore, although all examples so far deal with the direct and indirect effect of treatment on the same unit, this in no longer the case in this example as a human and an animal are different types of non-exchangeable units.

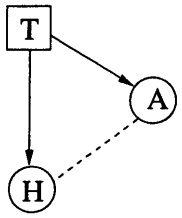

Figure C.1: The animal response A is a surrogate for the human response H. The edge between A and H is dashed because it is purely associational, and has no *causal* element.

Extending the direct indirect effects framework by making additional assumptions about the similarity between animals and humans could be an interesting avenue for further research

Bibliography

Angrist, J., G. Imbens, and D. Rubin (1996, June). Identification of Causal Effects using Istrumental Variables. *Journal of the American Statistical Association 91*(434), 444–455.

Cameron, S. and J. Heckman (1998). Life Cycle Schooling and Educational Selectivity: Models and Choice. *Journal of Political Economy 106*(2), 262–333.

Cooper, G. F. (1997). A Simple Constraint-Based Algorithm for Efficiently Mining Observational Databases for Causal Relationships. *Data Mining and Knowledge Discovery 1*, 203–224.

Dawid, A. and J. Dickey (1977). Likelihood and Bayesian inference from selectively reported data. *Journal of the American Statistical Association 72*(360), 845–850.

Dawid, A. P. (1979). Conditional Independence in Statistical Theory. *Journal of the Royal Statistical Society, Series B (Statisical Methodology) 41*(1), 1–31.

Dawid, A. P. (2000). Causal Inference without Counterfactuals (with

comments and rejoinder). *Journal of American Statistical Association 95*(450), 407–448.

Dawid, A. P. (2002). Influence diagrams for causal modelling and inference. *International Statistical Review 2*(70), 161–189.

Didelez, V. (2003). Graphical Models and Sequential Decisions. In *Proceedings of the 54th Session of the ISI.*

Freedman, D. and P. Humphreys (1999). Are there Algorithms that Discover Causal structure? *Synthese 121*, 29–54.

Greenland, S. (2005). Multiple-bias modelling for analysis of observational data. *Journal of the Royal Statistical Society, Series A (Statistics in Society) 2*(168), 1–25.

Heckerman, D., C. Meek, and G. Cooper (1999). A Bayesian approach to causal discovery. In C. Glymour and G. Cooper (Eds.), *Computation, Causation, and Discovery*, pp. 141–165. MIT Press, Cambridge, MA.

Heckerman, D. and R. Shachter (1995). Decision-Theoretic Foundations for Causal Reasoning. *Journal of Artificial Intelligence Research 3*, 405–430.

Heckman, J. and S. Navarro-Lozano (2004). Using matching, Instrumental Variables, and Control functions to estimate Economic Choice Models. *The Review of Economics and Statistics 80*(1), 30–57.

Heckman, J. and R. Robb (1985). Alternative Methods for Estimating the Impact of Interventions. In J. Heckman and B. Singer (Eds.), *Longitudi-*

nal Analysis of L*abor Market Data*, pp. 156–245. New York:Cambridge University Press.

Heckman, J. and E. Vytlacil (2005). Structural Equations,Treatment Effects and Econometric Policy Evaluation. *Econometrica 73*(3), 669–738.

Holland, P. W. (1986). Statistics and Causal Inference. *Journal of* A*merican Statistical Association 81*(396), 945–962.

Lauritzen, S. (1996). *Graphical* M*odels*. Clarendon Press, Oxford.

Lauritzen, S. (2001). Causal inference from graphical models. In O. Barndorff-Nielsen, D. R. Cox, and C. Klppelberg (Eds.), *Complex Stochastic Systems*, pp. 63–107. Chapman and Hall/CRC. London, Boca Raton.

Lauritzen, S. (2003). Graphical Models for Surrogates. In *Bulletin of the International Statitical Institute*, Volume 60, pp. 144–147.

Lauritzen, S. (2004). Discussion on Causality. *Scandinavian Journal of Statistics 31*(2), 189–201.

Lindley, D. (1985). *Making decisions* (2nd ed.). Wiley.

Lok, J. (2001). *Statistical Modelling of Causal Effects in Time*. Ph. D. thesis, Department of Mathematical Statistics, Free University of Amsterdam, The Netherlands.

MacDonald, J. and K. Smith (2004). The effects of technology-mediated

communication on industrial buyer behavior. *Industrial Marketing Management 33*, 107–116.

Neyman, J. (1923). On the Application of Probability Theory to Agricultural Experiments. Essay on Principles. Section 9, translated in. *Statistical Science 5*, 472–480.

Pearl, J. (1993). Comment: Graphical models, causality and intervention. *Statistical Science 8*, 266–269.

Pearl, J. (2000). *Causality.* Cambridge University Press.

Pearl, J. (2001a). Causal Inference in Health Studies: A Conceptual Introduction. *Health Services and Outcomes Research Methodology 2*(3-4), 189 – 220. Special issue on Causal Inference.

Pearl, J. (2001b). Direct and Indirect Effects. In M. Kaufmann (Ed.), *Proceedings of the 17th Conference on Uncertainty in Artificial Intelligence.*

Pearl, J. and T. Verma (1991). A theory of inferred causation. In J. Allen, R. Fikes, and E. Sandewall (Eds.), *Proceedings of Second International Conference on the Principles of Knowledge Representation and Reasoning*, pp. 441–452. Morgan Kaufmann.

Prentice, R. (1989). Surrogate endpoints in clinical trials. *Statistics in Medicine 66*, 431–440.

Raiffa, H. (1970). *Decision Analysis.* 2nd: Addison-Wesley.

Robins, J. (1995). Probabilistics Evaluation of Sequential Plans from Causal Models with Hidden Variables. In *Uncertainty in Artificial Intelligence, Proceedings of the 11th conference*, pp. 444–453.

Robins, J. (1998). Structural Nested Failure Time Models. In P. Armitage and T. Colton (Eds.), *Survival Analysis, P.K Andersen and N.Keiding, Section Editors. The Ecyclopedia of Biostatistics*, pp. 4372–4389. Chichester, UK, John Wiley and Sons.

Robins, J. and S. Greenland (1992). Identifiability and Exchangeability for Direct and Indirect Effects. *Epidemiology 3*, 143–155.

Robins, J. M. (1986). A New Approach to Causal Inference in Mortality Studies with a Sustained Exposure Period - Application to Control of the Healthy Worker Survivor Effect. *Mathematical Modelling 7*, 1393–1512.

Robins, J. M. (2003). Semantics of Causal DAG Models and the Identification of Direct and Indirect Effects. In N. Hjort, P. Green, and S. Richardson (Eds.), *Highly structured stochastic systems*, pp. 70–81. Oxford University Press.

Rosenbaum, P. and D. Rubin (1983). The Central Role of the Propensity Score in Observational Studies for Causal Effect. *Biometrika 70*(1), 41–55.

Rosenbaum, P. and D. Rubin (1985). Constructing a control group using multivariate matched sampling methods that incorporate the propen-

sity score. *American Statistician 39*, 35–39.

Rosenblatt, M. (1952). Remarks on a multivariate transformation. *The Annals of Mathematical Statistics 23*(3), 470–472.

Rubin, D. (2004). Direct and Indirect Causal Effects via potential outcomes. *Scandinavian Journal of Statistics 31*, 161–170.

Rubin, D. and J. Little (2002). *Statistical Analysis with missing data* (Second ed.). Wiley series in Probability and Statistics.

Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and non randomized studies. *Journal of Educational Psychology 66*(5), 699–701.

Rubin, D. B. (1978). Bayesian Inference for Causal Inference: The Role of Randomization. *Annals of Statistics 6*(1), 34–58.

Savage, L. (1954). *The Foundations of Statistics.* New York: Wiley.

Shafer, G. (1996). *The Art of Causal Conjecture.* MIT Press.

Silverstein, C., S. Brin, R. Motwani, and J. Ulman (2000). Scalable techniques for Mining Causal structures. *Data Mining and Knowledge Discovery 4*, 163–192.

Sober, E. (1988). Apportioning causal responsibility. *Journal of Philosophy 85*, 303–318.

Solanki, G., H. Schauffler, and L. S. Halpin Miller (2000). The Direct and Indirect Effects of Cost-Sharing on the Use of Preventive Services.

Health Services Research 34(6), 1331–1350.

Spirtes, P., C. Glymour, and R. Scheines (2000). *Causation, Prediction and Search.* New York, N.Y.: MIT Press.