Thesis for doctoral degree (Ph.D.) 2009

Causal Inference in Epidemiological Research

Arvid Sjölander

Karolinska Institutet

Thesis for doctoral degree (Ph.D.)
2009

Causal Inference in Epidemiological Research

Arvid Sjölander

Karolinska Institutet

From the Department of Medical Epidemiology and Biostatistics Karolinska Institutet, Stockholm, Sweden

Causal Inference in Epidemiological Research

Arvid Sjölander

Karolinska Institutet

Stockholm 2009

Published by Karolinska Institutet.
Printed by Larserics Digital Print AB.

ISBN 978-91-7409-277-6

Abstract

Traditionally, statistics has been viewed as the branch of science which deals with association. Many epidemiological research questions, however, are concerned with causation, not association. In this thesis we develop novel statistical methodology to address four epidemiological problems properly, from a causal inference point of view. We show, that for these four problems, our methods offer an attractive alternative to the 'standard' methodology, which may not yield the desired (causal) inference.

List of publications

This thesis is based on the following papers, which will be referred to in the text by their Roman numerals.

I. Sjölander A, Humphreys K, Palmgren J. (2008). On informative detection bias in screening studies. *Statistics in Medicine* **27**, 2635-2650.

II. Sjölander A, Humphreys K, Vansteelandt S, Bellocco R, Palmgren J. (2008). Sensitivity analysis for principal stratum direct effects, with an application to a study of physical activity and coronary heart disease. *Biometrics* DOI: 10.1111/j.1541-0420.2008.01108.x.

III. Sjölander A. (2008). Bounds on natural direct effects in the presence of confounded intermediate variables. *Statistics in Medicine* DOI: 10.1002/sim.3493.

IV. Sjölander A, Humphreys K, Vansteelandt S. (2008). A principal stratification approach to assess the differences in prognosis between cancers caused by hormone replacement therapy and by other factors. *Submitted*.

Contents

1 What is 'causal inference'?

Most epidemiological studies ultimately aim at detecting a causal relationship. Some examples of research questions which have been posed in epidemiology are:

1. Does long-term use of hormone replacement therapy (HRT) cause breast cancer (BC)?

2. To what extent is Alzheimer's disease caused by genetic or environmental factors?

3. Can obesity be prevented by moderate levels of physical activity?

One could therefore rightly claim that most epidemiological inference is, or at least strives to be, 'causal'. Nevertheless, there is a fundamental difference between how the concept of causality has been traditionally treated in epidemiology and biostatistics, and how it is treated in the modern framework of causal inference. Traditionally, causality has been defined, and inferred from observed associations, on an informal basis. During the last three decades, however, a formal theory of causal inference has been developed, with major contributions from Donald Rubin, James Robins, and Judea Pearl.

Consider research question 3 above. In an attempt to answer this question, a researcher might carry out a cross-sectional study and report the association between weight and physical activity. Suppose, for simplicity, that both weight (Y) and physical activity (X) are dichotomized, as '1' (obese/inactive) or '0' (not obese/active). One natural measure of the association between weight and physical activity is the relative risk

$$\beta = \frac{\mathrm{E}(Y|X=1)}{\mathrm{E}(Y|X=0)}. \tag{1}$$

If $\beta \neq 1$, then weight and physical activity are associated in the study population. This, however, does not imply that physical activity prevents (or, for $\beta < 1$, causes) obesity. If, for example, there are factors which both affect a subjects weight and physical activity level, then weight and physical activity may be associated even if activity does not prevent obesity; we say that β suffers from 'confounding', and we call the common causes of

weight and physical activity level 'confounders'[1]. The standard epidemiological way to deal with confounding is to carry out the analysis conditional on the confounders. The intuition behind this approach is that any discrepancy between exposed (physically active) and non exposed (inactive) in terms of the outcome (weight) must be attributed to the exposure, if all other relevant factors are held fixed. Formal causal theory supports the intuition in many scenarios. It can be shown, however, that conditioning on the confounders is not always a valid method. Furthermore, if the confounders are not measured, then conditioning is not possible in practice. A causal formalism is useful for a) determining when standard procedures fail to produce a valid (causal) inference, and b) to derive alternative methods for these situations.

As an illustration, suppose that the researcher addressing question 3 above carries out a longitudinal study instead of a cross-sectional study. For each study participant, measures of physical activity are obtained at different occasions, $t = 1, 2, ..., T$. We let X_t denote physical activity at time t, and define $\bar{X}_t \equiv \{X_1, X_2, ..., X_t\}$. Suppose that dietary intake is also measured at each occasion. We let L_t denote dietary intake at time t, and define $\bar{L}_t \equiv \{L_1, L_2, ..., L_t\}$. At the end of follow-up, weight (Y) is measured for each subject. By convention, we assume that L_t occurs before X_t. Thus, the temporal ordering of the variables is $\{L_1, X_1, L_2, X_2, ..., L_T, X_T, Y\}$. Suppose, for simplicity, that all variables are dichotomized. Given this setup, a standard way of assessing the association between weight and activity is to regress the mean of Y on activity history, $\bar{X}_T$. One possible regression model is

$$\text{logit E}(Y|\bar{X}_T) = \alpha + \beta \sum_{j=1}^{T} X_j. \tag{2}$$

Suppose that $\bar{L}_t$ is observed to be associated with both X_t and Y. Then we may suspect that $\bar{L}_t$ affects both X_t and Y, i.e. that $\bar{L}_t$ is a confounder for X_t and Y. Thus, we may be tempted to 'adjust' for dietary intake by adding $\bar{L}_T$ to the set of regressors, for example as

$$\text{logit E}(Y|\bar{X}_T, \bar{L}_T) = \alpha + \beta \sum_{j=1}^{T} X_j + \gamma \sum_{j=1}^{T} L_j. \tag{3}$$

[1]Although this is how many epidemiologists would define a 'confounder', this definition is not unproblematic, see Pearl (2000), Chapter 6, for a discussion

In model (2), the parameter β quantifies the marginal (over diet history) association between weight and physical activity history. In model (3), β quantifies the conditional (on diet history) association between weight and physical activity history. Does β in any of the models carry a causal interpretation? If so, in which? The answer, which probably appears counterintuitive to many epidemiologists, is that β may not have a causal interpretation in any of the models, *even if there is no unmeasured confounding*. More specifically, Robins (1986) showed that β may not have a causal interpretation in any of the models, even if there are no variables that affects both X_t and Y, at each occasion t, apart from $\bar{X}_{t-1}$ and $\bar{L}_t$. The example demonstrates the value of a formal approach to causal inference; using intuition alone it is often hard to determine whether a specific statistical analysis produces the desired causal inference. Moreover, without a formal approach it may not even be clear what the desired inference is. We will return to this example in Section 2. We will demonstrate how the the causal effect of activity on weight can be defined, why the standard analysis fails, and under what conditions the causal effect is identifiable.

The quest to formalize causality was initiated by Donald Rubin in the 70's, when he introduced the framework of *counterfactual variables*[2]. James Robins made important contributions to this framework in the 80's. During the 90's, Judea Pearl proposed a major generalization of this framework, based on non-parametric structural equations (NPSEs) and directed acyclic graphs (DAGs). In Section 2 we give a brief overview of these approaches. Using a formal causal framework it has been demonstrated that a number of 'standard solutions' are biased, i.e. do not produce a causal effect. The most well-known examples are probably the marginal and the conditional approaches in (2) and (3), examined by Robins (1986). Hence, the introduction of a formal framework has naturally led to a need for new statistical methodology. Examples of novel statistical methods, developed for making causal inference in situations where standard methods may fail, are propensity scores (Rosenbaum and Rubin, 1983), g-estimation (Robins, 1986), inverse probability weighting (Robins, 1997), principal stratification (Frangakis and Rubin, 2002), and instrumental variable techniques (Hernan and Robins, 2006).

[2]The main idea seems to have appeared originally in Rubin (1974), but the nomenclature was developed later.

A generic problem, to which special attention has been devoted in causal inference, is to make inference on a parameter (causal effect) which is not identifiable from data. Indeed, papers I, II, and III in this thesis deal with this problem. In Section 3 we briefly review the major approaches which have been proposed in the causal inference literature for dealing with unidentifiability.

Another common problem in causal inference is to estimate the direct effect of an exposure on an outcome. That is, the effect component which is not relayed by a specific intermediate variable. This problem is addressed in papers II and III. There are three common definitions of the direct effect in the literature. In Section 4 we review these definitions and discuss their interpretations.

Some of the definitions given below are formulated slightly more generally than in the cited papers, where they originally appear. In the text below I do not explicitly comment on the generalizations which are trivial (only on those which are not). For example, I claim in Section 4 that Pearl (2001) defined the controlled direct effect as some comparison of $\Pr\{Y(x', z)\}$, with $\Pr\{Y(x'', z)\}$. In fact, Pearl (2001) proposed the more restrictive definition $\mathrm{E}\{Y(x', z)\} - \mathrm{E}\{Y(x'', z)\}$.

2 The main causal frameworks

2.1 Counterfactuals

Suppose we want to learn about the causal effect of an exposure, X, on an outcome, Y. We let $X(u)$ and $Y(u)$ denote the exposure and outcome for subject u, respectively. Naturally, the exposure level varies from subject to subject. To define the causal effect of X on Y, however, we think of a hypothetical *intervention* which forces X to level x for each subject. We let $Y(x, u)$ denote the outcome Y for subject u under this intervention. The subject-specific outcomes $X(u)$, $Y(u)$, and $Y(x, u)$ are assumed to be deterministic functions of u[3]. We treat u as an outcome of a random variable U. Thus, the value of $X(U)$, $Y(U)$, and $Y(x, U)$ becomes random as well, abbreviated as X, Y, and $Y(x)$, respectively. Since the intervention which forces X to x is hypothetical, we say that $Y(x)$ is a *counterfactual* variable.

[3]More precisely, we *define* a 'subject' within this context to be a set of variables, or attributes, rich enough to render the outcomes $X(u)$, $Y(u)$, and $Y(x, u)$ deterministic.

We define the causal effect of taking X from x' to x'', as some comparison of the distribution $\Pr\{Y(x')\}$, with the distribution $\Pr\{Y(x'')\}$. Suppose, as in the weight-physical activity study from Section 1, that X and Y are binary, taking values 0 and 1. Then each subject u possesses two counterfactual outcomes, $Y(0,u)$, which is realized if the subject is forced to $X = 0$, and $Y(1,u)$, which is realized if the subject is forced to $X = 1$. We may for example define the causal relative risk as

$$\beta^* \equiv \frac{\mathrm{E}\{Y(1)\}}{\mathrm{E}\{Y(0)\}}. \tag{4}$$

Note that $\mathrm{E}\{Y(x)\}$ is an average over the whole population, under the intervention which sets X to x for each subject. Thus by comparing $\mathrm{E}\{Y(1)\}$ with $\mathrm{E}\{Y(0)\}$ we are comparing the same group of subjects, under two hypothetical interventions. In contrast, $\mathrm{E}(Y|X = x)$ is an average over the subset of the population for which X is observed to take value x. By comparing $\mathrm{E}(Y|X = 1)$ with $\mathrm{E}(Y|X = 0)$, as in the relative risk (1), we are comparing two different groups of subjects.

Since the intervention which forces X to x for each subject is hypothetical, the causal effect of X on Y is a hypothetical quantity as well. Nevertheless, we intuitively feel that it is sometimes possible to infer causal knowledge from observed associations, under reasonable assumptions. More specifically, we feel that randomization of the exposure should guarantee that an observed association can be interpreted causally. The following assumptions formalize this notion:

1. *Consistency*: $X(u) = x \Rightarrow Y(x,u) = Y(u)$.

2. *Weakly ignorable exposure assignment*: $Y(x) \amalg X \ \forall x$.

'Consistency' assures that if a subject attains levels $X = x$ and $Y = y$ in the absence of intervention, then the subject would also attain level $Y = y$ when forced to level $X = x$. Hence, the counterfactual outcome $Y(x)$ is observed and equal to Y whenever X is observed to take value x. 'Weak ignorability' assures that the distribution of counterfactual outcomes $Y(x)$ is the same within levels of X. To appreciate the idea, it is useful to consider $Y(x,u)$ to be an intrinsic characteristic of subject u, which is determined (although unobserved) before the subject-specific exposure level is determined. If the exposure is assigned randomly, then the characteristic $Y(x)$ should be independent of the exposure X. Weak ignorability thus follows naturally from

randomization of X[4]. Under consistency we have that

$$\Pr(Y = y|X = x) = \Pr\{Y(x) = y|X = x\}. \tag{5}$$

Under weak ignorability we have that

$$\Pr\{Y(x) = y|X = x\} = \Pr\{Y(x) = y\}. \tag{6}$$

Hence, from consistency and weak ignorability it follows that the distribution $\Pr\{Y(x)\}$ of counterfactual outcomes $Y(x)$ is observed, and equal to the distribution $\Pr(Y|X = x)$ of outcomes Y among those subjects who attained level $X = x$ in the absence of intervention. It follows, for example, that

$$\beta^* = \beta, \tag{7}$$

where β is the relative risk as defined in (1). We emphasize that whereas consistency is a rather weak assumption, and is routinely assumed in counterfactual analysis, weak ignorability is much stronger and often only reasonable within levels of covariates.

In papers II and IV we rely on the assumption of weak ignorability. To derive the results of paper III, however, it is necessary to assume strong ignorability.

A technical remark

Above, we defined $Y(x, u)$ as the outcome Y, for subject u, under the hypothetical intervention which forces X to x. We related this counterfactual variable to the factual variable Y through the consistency assumption. This definition is in line with Pearl (2000). There is a subtle difference, however, between this definition and how $Y(x, u)$ is usually defined in the papers by Donald Rubin (see for example Rubin (1974); Rosenbaum and Rubin (1983)). To understand the difference, consider the following quotation from Rosenbaum and Rubin (1983), page 41: 'We consider the case of two treatments, numbered 1 and 0. In principle, the ith of the N units under study has both a response r_{1i} that would have resulted if it had received treatment 1, and a response r_{0i} that would have resulted if it had received treatment 0. ... Since each unit receives only one treatment, either r_{1i} or r_{0i} is observed, but not

[4]Actually, physical randomization implies *strong* ignorability, defined as $Y(\cdot)\amalg X$, where $Y(\cdot)$ is the entire function (over x) of counterfactual outcomes. Strong ignorability implies weak ignorability, but not the other way around.

both ...'. In our notation, a 'unit' i is a 'subject' u, and the response r_{xi} is the counterfactual variable $Y(x, u)$. Rosenbaum and Rubin (1983) make no reference to interventions. Instead, they use the more vague formulation '.. had received treatment ...', without explicitly stating *how* the treatment had been received. Moreover, they appear to *define* $Y(x, u) \equiv Y$ when $X = x$, rather than formulating this as an assumption.

2.2 Non-parametric structural equations

The basic quantity in the counterfactual framework is the counterfactual variable. The counterfactual variable is used to define causal parameters of interest, and to formulate assumptions under which these parameters are identifiable. Typically, these assumptions are of the form $V_1 \amalg V_2 | V_3$, where V_1, V_2, and V_3 may be counterfactual variables, observed variables, or a combination of both. Weak ignorability is an example. The problem with counterfactual independence statements, such as weak ignorability, is that they are often hard to interpret. As a consequence, it is difficult to determine whether all relevant counterfactual independencies have been articulated, whether some of the articulated independencies are redundant, or even whether they are plausible and self-consistent.

As an illustration, consider the longitudinal weight-physical activity study from Section 1. Using counterfactual notation we let $Y(\bar{x})$ denote the random variable Y under the intervention which sets $\bar{X}$ to $\bar{x}$. We define the causal effect of taking $\bar{X}$ from $\bar{x}$ to $\bar{x}'$ as some comparison between $Pr\{Y(\bar{x})\}$ and $Pr\{Y(\bar{x}')\}$. Robins (1986) showed that this causal effect can be consistently estimated from observational data, if the following relation holds:

$$Y(\bar{x}) \amalg X_t | \bar{X}_{t-1}, \bar{L}_t \qquad \forall \bar{x}, t. \tag{8}$$

The relation in (8) is a time-dependent generalization of weak ignorability (see Section 2.1). Robins (1986) also demonstrated that even if (8) holds, neither marginalization over $\bar{L}_T$ (as in (2)), nor conditioning on $\bar{L}_T$ (as in 3), will in general produce the causal effect. In words, (8) states that the counterfactual weight level $Y(\bar{x})$, under the intervention which forces $\bar{X}$ to $\bar{x}$, is independent of the factual activity level X_t, conditional on activity history $\bar{X}_{t-1}$ and covariate history $\bar{L}_t$, for all occasions t. Many epidemiological researchers would probably have a hard time understanding the scientific meaning of the assumption in (8), let alone judging it's plausibility. The

reason why counterfactual independence statements are difficult to interpret, is that the human mind typically understands associations as by-products of causal mechanisms. For example, we explained the assumption of weak ignorability in Section 2.1 by referring to a specific assignment mechanism - *randomization*. When the independence statements are more complex, as in (8), it becomes increasingly difficult to mentally reconstruct an explaining mechanism, and thus, to interpret the statement.

Structural equations (SEs) represent an extension of the 'counterfactual approach'. In SEs, causal mechanisms are encoded explicitly. In this sense, SEs provide a more intuitive tool box for formulating causal assumptions than the counterfactual framework. Counterfactual variables can be derived from SEs, and independencies between counterfactual variables can be verified by simple visual (i.e. 'graphical') algorithms. Consider the equation

$$Y = \alpha + \beta X + \epsilon, \qquad \epsilon \sim N(0, 1). \tag{9}$$

The equation in (9) is a simple linear regression equation, which is commonly used in statistics. By labeling the equation as 'structural', however, we state that the left hand-side variable Y is not only associated with the right hand-side variable X, but *generated* by X, through a linear mechanism. The equality is assumed to hold irrespectively of how the value of X was generated. In particular, the equality is assumed to hold under intervention on X. This mechanistic interpretation of the regression equation stands in sharp contrast to the traditional 'associational' interpretation, under which X predicts Y in the absence of intervention, but may no longer be able to predict Y under intervention. In this sense a SE is asymmetric, and the equality sign in (9) behaves like the 'assignment equality', $:=$, commonly used in programming languages.

The equation in (9) can also be reformulated as

$$X = (Y - \alpha - \epsilon)/\beta. \tag{10}$$

This formulation gives the false impression that knowing Y and ϵ (and (α, β)) renders X known. This is only true in the absence of intervention. If we intervene and force Y to, say y, then the value of X can no longer be read off from (9), since X causes Y, but not the other way around.

In many causal inference applications, we want to express *which* explanatory variables affect (generate) a certain outcome, without specifying *how* the

mechanism behaves (i.e. whether the correct structural model is linear, log-linear, non-linear etc). It is for this purpose that non-parametric structural equations (NPSEs) are used. The non-parametric version of the structural equation in (9) is given by

$$Y = F(X, \epsilon), \tag{11}$$

where $F(\cdot)$ is an unspecified function, and the error term ϵ follows an unspecified distribution.

We often encounter problems in which several variables appear, which may be related to each other through a complex scheme of causal mechanisms. To express this scheme, a NPSE system is used. Formally, a NPSE system is a triple $(\mathbf{U}, \mathbf{V}, \mathbf{F})$, in which $\mathbf{U}$ is a set of exogenous variables, following an unspecified joint distribution, $\mathbf{V}$ is a set of endogenous variables (disjoint from $\mathbf{U}$), and $\mathbf{F} = \{F_V(\mathbf{R}_V)|V \in \mathbf{V}\}$, where $F_V(\mathbf{R}_V)$ is a function that deterministically assigns a value to V for each setting of the remaining variables $\mathbf{R}_V = \mathbf{U} \cup \mathbf{V} \setminus V$. The words 'endogenous' should be interpreted as 'explained by the model'. Similarly, 'exogenous' means 'not explained by the model'. Without loss of generality, the exogenous variables can be assumed to have no common causes, which implies that they are independent. As an example, consider the NPSE system

$$\begin{array}{rcl} X & = & F_X(U_X) \\ Y & = & F_Y(X, U_Y) \end{array} \tag{12}$$

For this example, $\mathbf{U} = \{U_X, U_Y\}$, $\mathbf{V} = \{X, Y\}$, and $\mathbf{F} = \{F_X(U_X), F_Y(X, U_Y)\}$.

A NPSE system is assumed to be *autonomous*, which means that an intervention which forces a variable V to a certain value, v say, and thus overrides the function $F_V(\mathbf{R}_V)$, leaves the remaining functions unaltered. Forcing X to x thus modifies the NPSE system in (12) as

$$\begin{array}{rcl} X & = & x \\ Y & = & F_Y(X, U_Y) \end{array} \tag{13}$$

Forcing Y to y modifies the system as

$$\begin{array}{rcl} X & = & F_X(U_X) \\ Y & = & y \end{array} \tag{14}$$

The connection between counterfactual variables and NPSE systems is clear if we think of one particular realization, u, of the exogenous variables

$\mathbf{U}$ as representing a 'subject' in the counterfactual framework. For the system in (12), the counterfactual outcome $Y(x,u)$ is obtained by forcing X to x, and is thus equal to $F_Y(x,u_Y)$. In contrast, the *factual* outcome $Y(u)$ is obtained by letting X attain its factual value of $X(u) = F_X(u_X)$, and is thus equal to $F_Y(X(u),u_Y) = F_Y(F_X(u_X),u_Y)$. The random variables $Y(x)$ and Y are obtained by considering $\mathbf{U} = \{U_X, U_Y\}$ as random; $Y(x) = F_Y(x,U_Y)$ and $Y = F_Y(F_X(U_X),U_Y)$.

We demonstrated in Section 2.1, that the causal effect of X on Y, defined as some comparison of $\Pr\{Y(1)\}$ against $\Pr\{Y(0)\}$, is identifiable in the absence of intervention on X, if both consistency and weak ignorability hold. Assume that we want to investigate whether the causal effect of X on Y is identifiable in system (12), in the absence of intervention. Consistency of counterfactuals is implied in NPSE systems by the assumption of autonomy. Thus, it remains to investigate whether weak ignorability holds for the system in (12). We can do this algebraically as follows. From the previous paragraph we have that $X = F_X(U_X)$ and $Y(x) = F_Y(x,U_Y)$. By convention, $U_X \amalg U_Y$. Hence, for the structural system in (12), $Y(x) \amalg X$, which implies that $\Pr\{Y(x)\}$ is identifiable and equal to $\Pr(Y|X=x)$. This example highlights that the system in (12) encodes an important structural assumption, namely that X and Y do not have any common causes ($\mathbf{R}_X \cup \mathbf{R}_Y = \emptyset$). The immediate implication of this assumption is that the causal effect of X on Y is identifiable in the absence of interventions. If we find the assumption of no common causes implausible, we may add a common cause, U_{XY}, of X and Y, by modifying the system in (12) as

$$\begin{array}{rcl} X & = & F_X(U_X, U_{XY}) \\ Y & = & F_Y(X, U_X, U_{XY}) \end{array} \tag{15}$$

For the system in (15) we have that $X = F_X(U_X,U_{XY})$ and $Y(x) = F_Y(x,U_Y,U_{XY})$. Since the random variable U_{XY} appears in both the expression for X and $Y(x)$, it follows that X and $Y(x)$ are dependent, and weak ignorability does not hold.

2.3 Directed acyclic graphs

A NPSE system contains all relevant information about the causal relations between the variables in $\mathbf{U} \cup \mathbf{V}$. This information, however, is captured in a rather dense form. An expository way of presenting a NPSE system, is

through a directed acyclic graph (DAG). For a comprehensive introduction to DAGs, see Pearl (2000). On a DAG, each variable in $\mathbf{U} \cup \mathbf{V}$ is represented by a node, and each direct causal influence is represented by a directed edge. More specifically, we draw an arrow from X to Y iff $F_Y(\mathbf{R}_Y)$ is non-trivial in X. By convention, exogenous variables which only appear in a single equation in the NPSE system are not displayed on the DAG. For example, the NPSE system in (12) is represented by the DAG in Figure 1. The NPSE system in (15) is represented by the DAG in Figure 2.

Figure 1: Graphical representation of the NPSE system in (12).

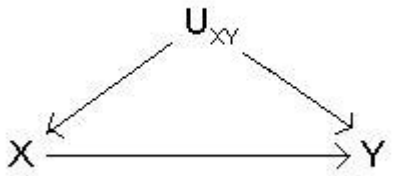

Figure 2: Graphical representation of the NPSE system in (15).

Often, we want to investigate whether two variables, V_1 and V_2, on a DAG, are independent, possibly conditional on a third (set of) variable(s), V_3. Pearl (2000) presented a convenient algorithm for this task (the algorithm originally appeared in Verma and Pearl (1988)). The following definition is from Pearl (2000):

A path[5] *p is said to be blocked by a set of nodes* V_3 *if and only if*

1. *p contains a chain* $i \to m \to j$ *or a fork* $i \leftarrow m \to j$ *such that the middle node* m *is in* V_3*, or*

[5]The word 'path' refers to any unbroken, non-intersecting route along the edges of a DAG, which may go either along or against the arrows.

2. *p contains an inverted fork* $i \rightarrow m \leftarrow j$ *such that the middle node* m *is not in* V_3 *and such that no descendant*[6] *of* m *is in* V_3.

A set V_3 *is said to d-separate nodes* V_1 *and* V_2 *if and only if* V_3 *blocks every path from* V_1 *to* V_2.

According to Pearl (2000), V_1 and V_2 are independent, given V_3, if V_3 d-separates V_1 and V_2. Conversely, if V_3 doesn't d-separate V_1 and V_2, then V_1 and V_2 are in general associated, given V_3.

To appreciate the usefulness of d-separation, consider the longitudinal weight-physical activity study from Section 1. Assume that $T = 2$, and that the variables X_T, L_T, and Y are causally related through the NPSE system depicted in Figure 3. Under the structural model in Figure 3, there is no

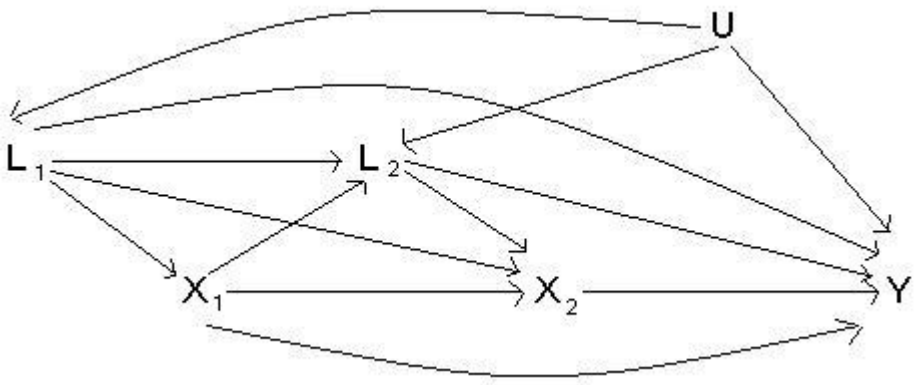

Figure 3: A possible graphical representation of the longitudinal weight-physical activity study.

unmeasured confounding in the sense that there are no variables that affect both X_t and Y, at each occasion t, apart from $\bar{X}_{t-1}$ and $\bar{L}_t$. Nevertheless, we can use d-separation to prove that neither marginalization over $\bar{L}_T$ (as in (2)), nor conditioning on $\bar{L}_T$ (as in (3)), will in general produce the causal effect of $\bar{X}_T$ on $\bar{Y}_T$. To do this, we proceed in two steps. First we modify the DAG in Figure 3 so that it represents a possible null hypothesis of no causal effect of physical activity on weight. Then we use d-separation to show that, for this null hypothesis, $\bar{X}_T$ is associated with Y both marginally, and conditionally

[6]A node k is said to be a 'descendant' of m if k can be reached by following the arrows from m.

on $\bar{L}_T$. Hence, as shown by Robins (1986), neither marginalization over $\bar{L}_T$ (as in (2)), nor conditioning on $\bar{L}_T$ (as in 3), will in general produce the zero causal effect (i.e. $\hat{\beta} \neq 0$). In the first step we modify the DAG in Figure 3 by removing all arrows into Y, except from U. The resulting DAG is displayed in Figure 4. In Figure 4, there is no directed path from X_t to Y, for any t.

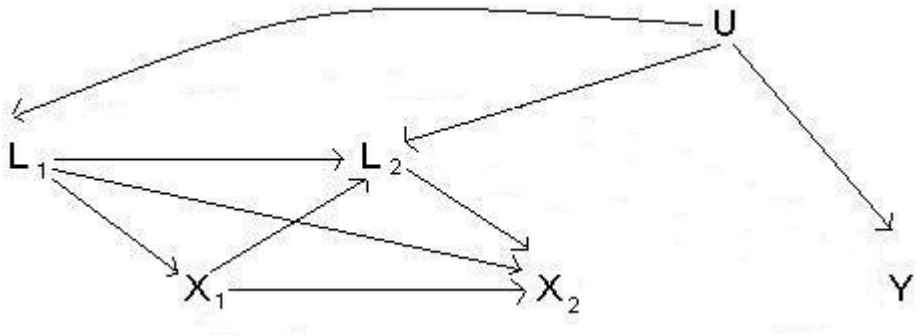

Figure 4: A possible graphical representation of the null hypothesis of no causal effect of physical activity on weight.

Thus, the DAG in Figure 4 represents a possible null hypothesis of no causal effect of $\bar{X}_T$ on Y. By applying d-separation, however, we observe that $\bar{X}_T$ is associated with Y marginally. This is because the paths $X_1 \leftarrow L_1 \leftarrow U \rightarrow Y$ and $X_2 \leftarrow L_2 \leftarrow U \rightarrow Y$ are not blocked, according to the two rules above. We also observe that $\bar{X}_T$ is associated with Y conditionally on $\bar{L}_T$. This is because the path $X_1 \rightarrow L_2 \leftarrow U \rightarrow Y$ is not blocked in Figure 4 (according to rule 2 it would have been blocked had we not conditioned on L_2, but then the path $X_2 \leftarrow L_2 \leftarrow U \rightarrow Y$ would have been unblocked).

Pearl (2000) also demonstrates (Section 7.1.4) that the method of d-separation can be extended to evaluate counterfactual independence statements of arbitrary complexity. Using this extension, it is an easy task to show that the statement in (8) holds for the DAG in Figure 3. Hence, the causal effect of activity on weight is identifiable, under this structural model, but can not be estimated with standard methods[7].

The example illustrates that once a causal model has been formulated in terms of causal mechanisms through a NPSE system/DAG, then complex

[7]Robins (1986) derived an expression for this causal effect, as a functional of the joint distribution of $(\bar{X}_T, \bar{L}_T, Y)$. He called it *the G-functional.*

statements such as (8) can be easily evaluated in a straightforward algorithmic fashion.

3 Identifiability, bounds and sensitivity analysis

3.1 Identifiability

Many problems in causal inference are concerned with making inference on a parameter which is unidentifiable. This is indeed the case for papers I-III in this thesis. To explain what we mean by 'unidentifiable', consider as a motivating (toy) example a study in which the aim to estimate the effect of a binary exposure X on a binary outcome Y. Specifically, the aim is to estimate the causal relative risk β^* as defined in (4). A fixed number of exposed subjects, and a fixed number of unexposed subjects, are enrolled. Suppose that the outcome is only measured for a subset of the study participants, for the remaining subjects the outcome is missing. Let M be a missing outcome indicator, i.e. $M = 1$ if the outcome is missing, and $M = 0$ if the outcome is not missing. Suppose, for simplicity, that the analysis is restricted to the complete data, i.e. subjects with $M = 1$ are 'thrown away'. The data thus consists of one iid sample from $\Pr(y|X = 1, M = 0)$, and one iid sample from $\Pr(y|X = 0, M = 0)$. We define

$$\begin{aligned} p_x &\equiv \Pr(Y = 1|X = x, M = 0), \\ \mathbf{p} &\equiv (p_0, p_1), \\ \alpha &\equiv \Pr(Y = 1|X = 0), \\ \eta &\equiv \frac{\Pr(M = 0|Y = 0)}{\Pr(M = 0|Y = 1)}. \end{aligned}$$

Suppose that we are willing to assume that the causal mechanisms relating X, Y and M are as in Figure 5. The DAG in Figure 5 encodes the important assumption that X does not have any influence on M, other than through Y. In addition, it implies that (X, Y, M) have no unmeasured common causes. It follows, that the causal relative risk β^* is identical to the relative risk β, as defined in (1). Unfortunately, since missingness depends on the outcome,

Figure 5: A possible causal mechanisms for X, Y, and M.

β is unidentifiable. To see this, we use Bayes rule to rewrite $\mathbf{p}$ as

$$\begin{aligned} p_1 &= \frac{\alpha\beta}{\alpha\beta + \eta(1-\alpha\beta)}, \\ p_0 &= \frac{\alpha}{\alpha + \eta(1-\alpha)}. \end{aligned} \tag{16}$$

The right hand-side of the system in (16) consists of three free parameters, and the equation system can not be solved to yield a unique solution for β. This implies that even if we knew the observed data distribution, $\mathbf{p}$, we could still not infer the true value of β. We say that β is *unidentifiable*.

Generally, 'identifiability' can be defined as follows (Greenland, 1999):

Consider a vector $\mathbf{Z}$ *of random variables having a distribution* $F(\mathbf{z})$ *that depends on an unknown parameter vector* θ. θ *is* identifiable *by observation of* $\mathbf{Z}$ *if distinct values for* θ *yield distinct distributions for* $F(\mathbf{z})$.

In other words; θ is identifiable by observations $\mathbf{Z}$ if θ is a function(al) of $F(\mathbf{z})$:

$$\theta = g\{F(\mathbf{z})\}.$$

3.2 Bounds

That a parameter is unidentifiable does not mean that data carries no information about the parameter. For example, careful investigation of the equation system in (16) shows that although it has no unique solution for β, not every value $\beta \in (0, \infty)$ is a feasible solution. On the contrary, the feasible solutions are given by

$$\beta \in \left\{ \min\left(1, \frac{p_1(1-p_0)}{p_0(1-p_1)}\right), \max\left(1, \frac{p_1(1-p_0)}{p_0(1-p_1)}\right) \right\} \tag{17}$$

Hence, β is *bounded* by the observed data distribution $\mathbf{p}$. In practice, the true value of $\mathbf{p}$ is unknown. Estimated bounds for β can be obtained by replacing $\mathbf{p}$ in (17) by an estimate, $\hat{\mathbf{p}}$. To acknowledge the additional uncertainty in β due to sampling variability, we may for example report standard errors for the upper and lower bound.

In general, we say that θ and $F(\mathbf{z})$ are *variation independent* if the joint space for θ and $F(\mathbf{z})$ is the cartesian product of the individual spaces for θ and $F(\mathbf{z})$. That θ and $F(\mathbf{z})$ are *not* variation independent implies that not all possible values of β are compatible with each specific distribution $F(\mathbf{z})$. Thus, if θ and $F(\mathbf{z})$ are not variation independent, then bounds for θ can be constructed, as a function of $F(\mathbf{z})$. In practice, $F(\mathbf{z})$ is unknown and the bounds for θ have to be estimated from data. When Z is high-dimensional and/or sparse, the practical usefulness of the bounds is often very limited because of a large sampling variability.

3.3 Sensitivity analysis

In many practical situations we have prior knowledge which can be used to sharpen our conclusions about an unidentifiable parameter. To continue the motivating example, suppose that we believe that subjects who experience the outcome ($Y = 1$) are more likely to be missing than subjects who don't ($Y = 0$). In terms of the model parameters, this assumption translates to $\eta > 1$. In addition, we may find it implausible that subjects with $Y = 1$ are more than, say, twice as likely to be missing, as subjects with $Y = 0$. Thus, we believe that $1 < \eta < 2$. To see how this prior knowledge translates into conclusions about β, we may use the system in (16) to express β as a function(al) of (p_0, p_1), and η:

$$\beta = \frac{p_1\{1 + p_0(\eta - 1)\}}{p_0\{1 + p_1(\eta - 1)\}}.$$

Hence, for given values of $\mathbf{p}$, each value of η maps to one single value for β. If $\eta = 1$, then $\beta = p_1/p_0$; we say that the missingness is 'ignorable'. If $\eta \neq 1$, then $\beta \neq p_1/p_0$ and we say that the missingness is 'nonignorable'. By varying η over the plausible range $(1, 2)$, we obtain a range of plausible values for β. We say that we perform a *sensitivity (to nonignorable missingness) analysis* for β. We call η a *sensitivity analysis parameter*. Replacing $\mathbf{p}$ with an estimate yields an estimated range of plausible values for β. Rotnitzky

et al (1998) proposed a method of sensitivity (to nonignorable missingness) analysis, which is applicable to longitudinal data.

In general, to perform a sensitivity analysis for an unidentifiable parameter θ, we must find a parameter η such that, for each fixed value of η, θ is identifiable (by observations Z). In other words, we must require that

$$\theta = g\{F(\mathbf{z}), \eta\}. \tag{18}$$

If $F(\mathbf{z})$ is known, we can use the relation in (18) to map each plausible value of η into a value for θ. In practice, $F(\mathbf{z})$ is unknown and θ has to be estimated for each value of η. For low-dimensional problems, such as the motivating example above, we may replace the population distribution $F(\mathbf{z})$ in (18) by the sample distribution $\hat{F}(\mathbf{z})$ to yield a non-parametric estimator for θ. For high-dimensional problems, additional parametric assumptions are often required[8].

In more complicated scenarios, the challenge in carrying out a successful sensitivity analysis lies in the choice of sensitivity analysis parameter. A minimal criterion for a parameter η to be able to function as a sensitivity analysis parameter for θ, is that (18) holds. In addition, it is important that η is easy to interpret. If η is difficult to interpret, then it may be hard, even for a subject matter expert, to specify a range of plausible values for η. Finally, it is desirable that η is variation independent of $F(\mathbf{z})$. If $F(\mathbf{z})$ is known, and η and $F(\mathbf{z})$ are variation dependent, then we must restrict a sensitivity analysis to the set of values for η which are compatible with $F(\mathbf{z})$. To find this set is sometimes non-trivial, which makes a sensitivity analysis difficult. In most practical scenarios, we have a sample from $F(\mathbf{z})$, but the distribution itself is unknown. This makes the problem of variation dependence more subtle. The sample obviously carries information about $F(\mathbf{z})$. Hence, if η and $F(\mathbf{z})$ are variation dependent, the sample should, in some sense, also carry information about η. To extract this information from data, however, is often not a trivial task. Thus, we may fail to notice when η is taken to values which are not consistent with data. This problem does not arise if η and $F(\mathbf{z})$ are variation independent.

[8]When θ has to be estimated, the relation in (18) is not often used explicitly, Typically, an estimate of θ is obtained, for each value of η, as the solution to an unbiased estimating equation in (θ, η), for example the ML score equation.

4 Direct effects

4.1 Definitions

A common goal of epidemiological research is to separate direct effects from indirect effects. As an illustration, consider the DAG in Figure 6. In Figure

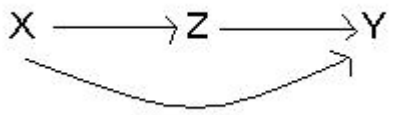

Figure 6: A graphical illustration of direct and indirect effects.

6, X affects Y directly, but also indirectly through the intermediate variable Z. Suppose we are interested in the direct effect of X on Y. There are three common definitions of direct effects in the literature; *controlled*, *natural*, and *principal stratum* direct effects. In this section I review these definitions, and discuss their interpretations.

The most intuitive way to measure the direct effect of X on Y, is to hold Z fixed by an intervention. The controlled direct effect (CDE) of taking X from x' to x'', on Y, at level $Z = z$, is defined as some comparison of $\Pr\{Y(x', z)\}$, with $\Pr\{Y(x'', z)\}$. For binary X, we may for example define the controlled direct mean ratio, at $Z = z$, as

$$\beta_z = \frac{\mathrm{E}\{Y(1, z)\}}{\mathrm{E}\{Y(0, z)\}}.$$

The CDE is intuitively compelling. In some scenarios, however, it may not answer the scientific question of interest. For illustration, consider the following example, borrowed from Pearl (2001). Suppose that X represents the intake of a medical drug ($X = 1$ for 'intake', $X = 0$ for 'no intake') and Y represents a binary disease outcome ($Y = 1$ for 'disease', $Y = 0$ for 'no disease'). Suppose further that the drug has a side-effect; it causes headache. People who suffer from headache tend to take aspirin ($Z = 1$ for 'aspirin intake', $Z = 0$ for 'no aspirin intake'), which, in turn may have its own effect on the disease. Suppose that the drug manufacturer is considering ways of eliminating the side-effect. A natural question to ask is whether the

drug would retain its effectiveness when the side-effect is eliminated. The CDE does not answer this question, since it refers to a specific aspirin level, taken uniformly by all subjects. The target population is one where aspirin intake varies from individual to individual, depending on factors beside drug-induced headache. As a meaningful measure of effectiveness of the drug for the target population, Pearl (2000) suggested to use the natural direct effect (NDE). Let $Z(x,u)$ denote the outcome Z for subject u, under the hypothetical intervention which sets X to x. Pearl (2000) defined the NDE of taking X from x' to x'', on Y, at level $X=x$, as some comparison of $\Pr[Y\{x', Z(x)\}]$, with $\Pr[Y\{x'', Z(x)\}]$. For binary X, we may for example define the natural direct mean ratio, at $X=x$, as

$$\gamma_x = \frac{\mathrm{E}[Y\{1, Z(x)\}]}{\mathrm{E}[Y\{0, Z(x)\}]}.$$

If we let $X=0$ correspond to 'no drug', then γ_0 measures the effectiveness of the drug for a population where everybody attains the same level of aspirin as they would, had they not used the drug.

We may formulate the CDE and the NDE in terms of interventions in a NPSE system. The DAG in Figure 6 represents the system

$$\begin{array}{rcl} X &=& F_X(U_X) \\ Z &=& F_Z(X, U_Z) \\ Y &=& F_Y(X, Z, U_Y) \end{array} \tag{19}$$

The CDE refers to a scenario where the system in (19) is modified as

$$\begin{array}{rcl} X &=& x \\ Z &=& z \\ Y &=& F_Y(X, Z, U_Y) \end{array} \tag{20}$$

The CDE is obtained by forcing Z to z in the second equation of (20), and observing how Y varies in the last equation, when X is varied in the first equation. The natural direct effect effect refers to a scenario where the system in (19) is modified as

$$\begin{array}{rcl} X &=& x' \\ Z &=& F_Z(x, U_Z) \\ Y &=& F_Y(X, Z, U_Y) \end{array} \tag{21}$$

The NDE is obtained by forcing Z to $F_Z(x, U_Z)$ in the second equation of (21), and observing how Y varies in the last equation, when X is varied in the first

equation. The NPSE formulation highlights an important difference between the CDE and the NDE. The CDE can be directly measured in an experiment where X and Z are controlled and can be forced to fixed values. To directly measure the NDE, however, we must be able to carry out an experiment where Z can be forced to follow a particular distribution. For this reason, Robins (2003) referred to the NDE as a 'non-manipulative' parameter. In some special cases, however, the NDE is indeed 'manipulative', in the sense of Robins (2003). Consider the aspirin-example borrowed from Pearl (2001). Suppose that the drug can be separated into two components, X_1 and X_2, where X_1 is the component which causes head-ache, and X_2 is the component which directly influences Y. More specifically, we assume that

$$\begin{array}{rcl} X & = & F_X(U_X) \\ X_1 & = & X \\ X_2 & = & X \\ Z & = & F_Z(X_1, U_Z) \\ Y & = & F_Y(X_2, Z, U_Y) \end{array} \tag{22}$$

The deterministic relation between X and (X_1, X_2) reflects the fact that the two components are present ($X_1 = X_2 = 1$) if and only if the drug is taken ($X = 1$). For the system in (22), the NDE of X on Y, at $X = x$, is a manipulative parameter, since it can be obtained by holding X_1 fixed at x, and varying X_2.

The CDE and the NDE may be of scientific interest if the intermediate variable, Z, could, at least hypothetically, be manipulated. In some scenarios, even hypothetical interventions on Z may be hard to imagine. As an example, suppose that $Z = 1$ represents being alive 1 year after treatment (X) assignment, and Y represents an outcome measured 1 year after assignment. Obviously, Y is only defined if $Z = 1$. To measure the effect of X on Y we might want to keep each subject alive until Y is measured. Doing so, we would obtain the CDE of X on Y, at $Z = 1$. Preventing death, however, is a highly hypothetical intervention, which raises the question of whether the CDE is a meaningful quantity in this context. Rubin (2004) offered an alternative definition of a direct effect, based on *principal stratification* (Frangakis and Rubin, 2002), which does not rely on intervention on Z. We say that a subject, u, belongs to principal stratum z if $Z(x, u) = z \ \forall x$. Note that in most realistic scenarios $Z(x, u)$ is only observed for $x = X(u)$ (see Section 2.1). Thus, a principal stratification is a hypothetical partitioning of the population, which is often not possible to carry out in practice. For subjects

within the principal stratum $\{Z(x) = z\ \forall x\}$, a change in X never results in a change in Z. Thus, if we observe that a change in X results in a change in Y, for a subject within $\{Z(x) = z\ \forall x\}$, we may conclude that, for this subject, X had a direct effect on Y. Rubin (2004) defined the principal stratum direct effect (PSDE) of taking X from x' to x'', on Y, for $\{Z(x) = z\ \forall x\}$, as some comparison of $\Pr\{Y(x')|Z(x) = z\ \forall x\}$, with $\Pr\{Y(x'')|Z(x) = z\ \forall x\}$. For binary X, we may for example define the principal stratum direct mean ratio, at $Z = z$, as

$$\delta_z = \frac{\mathrm{E}\{Y(1)|Z(x) = z\ \forall x\}}{\mathrm{E}\{Y(0)|Z(x) = z\ \forall x\}}.$$

When $Z = 1$ represents 'being alive 1 year after assignment', δ_1 represents the PSDE of X on Y, for those who stay alive regardless of whether they are assigned to $X = 0$ or $X = 1$.

Since the definition of the PSDE does not require Z to be manipulable, the PSDE is applicable to a wider range of problems than the CDE and the NDE. On the other hand, it is not always correct to interpret the PSDE as a direct effect. In this sense, the term principal stratum *direct effect* is a misnomer. To see this, consider the following example, which is adapted from Robins et al (2007). Suppose that we don't observe Z directly. Instead we observe a coarse version of Z, defined as $Z^* = F_{Z^*}(Z)$. We would then perhaps be interested in the effect component of X on Y which is not relayed through Z^*. We may for example seek to estimate the CDE of taking X from x' to x'', on Y, at $Z^* = z^*$, defined as some comparison of $\Pr\{Y(x', z^*)\}$, with $\Pr\{Y(x'', z^*)\}$. But, since Z^* is only a coarse measurement of Z, it is not clear what we mean by 'a hypothetical intervention which sets Z^* to z^*'. Hence, $Y(x, z^*)$ is not well defined. Nothing prevents us, however, from conditioning on the event $Z^*(x) = z^*\ \forall x \Leftrightarrow F_{Z^*}(Z) = z^*\ \forall x$. Hence, the PSDE of taking X from x' to x'', on Y, for $\{Z^*(x) = z^*\ \forall x\}$, is well defined as some comparison of $\Pr\{Y(x')|Z^*(x) = z^*\ \forall x\}$, with $\Pr\{Y(x'')|Z^*(x) = z^*\ \forall x\}$. On the other hand, for this example a PSDE of X on Y is not necessarily 'direct'. Suppose, for example, that

$$\begin{aligned} Z &= X + U_Z && (23) \\ Z^* &= \begin{cases} 0 & \text{if} \quad Z = 0 \\ 1 & \text{if} \quad Z \neq 0 \end{cases} \\ Y &= Z, && (24) \end{aligned}$$

and that X and U_Z are binary, taking values 0 and 1. Since X does not appear in the equation for Y, X has no direct effect on Y. Calculating the PSDE at $\{Z^*(x) = 1 \ \forall x\}$, however, gives that $\mathrm{E}\{Y(1)|Z^*(x) = 1 \ \forall x\} = 2$ and $\mathrm{E}\{Y(0)|Z^*(x) = 1 \ \forall x\} = 1$. The reason for this, is that X has an effect on Y, which is mediated through Z. This 'indirect effect' appears as a PSDE, since Z^* lumps together several values of Z. Informally, we may say that Z^* fails to completely block the path $X \to Z \to Y$.

4.2 Identification

For the DAG in Figure 6 (that is, when $\{X, Z, Y\}$ have no common causes), the CDE, the NDE, and the PSDE are all identifiable. To prove this we need the following results, which all holds for the DAG in Figure 6:

$$Z(x) \amalg X \qquad \forall x, \tag{25}$$

$$Y(x, z) \amalg (X, Z) \qquad \forall x, z, \tag{26}$$

$$Y(x', z) \amalg Z(x) \qquad \forall x', z, x. \tag{27}$$

The CDE is identifiable because

$$\begin{aligned} \Pr\{Y(x, z) = y\} &= \Pr\{Y(x, z) = y|X = x, Z = z\} \\ &= \Pr(Y = y|X = x, Z = z), \end{aligned} \tag{28}$$

were the first equality follows from (26), and the second from consistency of counterfactuals. The NDE is identifiable because

$$\begin{aligned} \Pr[Y\{x', Z(x)\} = y] &= \int_z \Pr\{Y(x', z) = y|Z(x) = z\}\Pr\{Z(x) = z\}dz \\ &= \int_z \Pr\{Y(x', z) = y|X = x', Z = z\}\Pr\{Z(x) = z|X = x\}dz \\ &= \int_z \Pr\{Y = y|X = x', Z = z\}\Pr(Z = z|X = x)dz, \end{aligned}$$

where the first equality follows from consistency and standard probability rules, the second from (25)-(27), and the third from consistency. The PSDE

is identifiable because

$$\begin{aligned}\Pr\{Y(x') = y | Z(x) = z \ \forall x\} &= \Pr\{Y(x', z) = y | Z(x) = z \ \forall x\} \\ &= \Pr\{Y(x', z) = y\} \\ &= \Pr(Y = y | X = x', Z = z),\end{aligned}$$

where the first equality follows from consistency, the second from (27), and the third from (28).

In more complicated scenarios, identification of direct effects is often non-trivial. In paper II and III we consider a scenario in which Z and Y have unmeasured common causes, as displayed by the DAG in Figure 7. For the

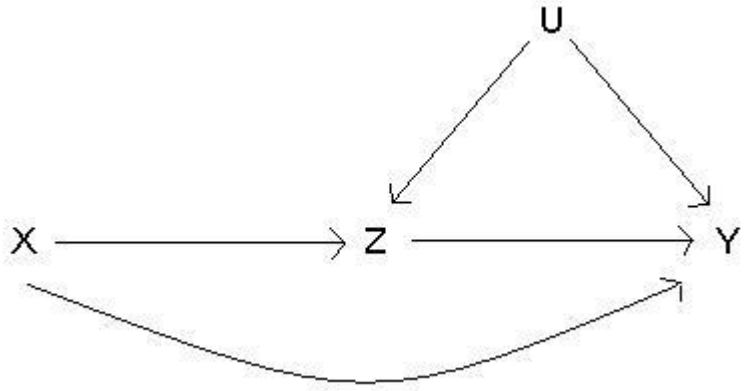

Figure 7: The DAG associated with papers II and III.

DAG in Figure 7, (25)-(27) do not hold and none of the three direct effects is identifiable.

5 Summary of the papers

5.1 Paper I

In paper I we consider estimation of the effect of hormone replacement therapy (HRT) on breast cancer. It is well known that women who use HRT are more likely to use screening mammography, than women who do not use HRT. Hence, existing tumors are also more likely to be detected in the

treated arm, even if HRT does not cause cancer. A naive way to adjust for this detection bias is to condition on screening behavior. We use DAGs to formalize the problem. Using d-separation we demonstrate that the naive analysis will not produce the causal HRT effect, ψ (defined in equation (2) of paper I), if screening behavior and disease status have unmeasured common causes. Under this scenario, ψ is not identifiable. We show that ψ is not variation independent of the observed data distribution. We use this result to construct bounds for ψ. We show how a sensitivity analysis for ψ can be carried out, using the relative screening efficiency, η (defined in equation (4) of paper I), as a sensitivity analysis parameter. η is not variation independent of the observed data distribution, which must be taken into account in this sensitivity analysis.

Joffe and Colditz (1998) proposed a method of dealing with detection bias in screening studies, based on 'data restriction'. We show that the method of data restriction is not valid.

A technical remark

As stated in the paper (page 2645), the method of data restriction is based on the erroneous decomposition

$$\mathrm{E}\{Y(x)\} = \sum_z \Pr(Z = z|X = x)\mathrm{E}\{Y(x, z)\}. \tag{29}$$

Equation (27) in the paper suggests one correct decomposition. To understand why the decomposition in (29) is erroneous, it is instructive to consider the alternative (correct) decomposition

$$\begin{aligned} \mathrm{E}\{Y(x)\} &= \sum_z \Pr\{Z(x) = z\}\mathrm{E}\{Y(x, z)|Z(x) = z\} \\ &= \sum_z \Pr(Z = z|X = x)\mathrm{E}\{Y(x, z)|Z(x) = z\}. \end{aligned} \tag{30}$$

The second equality in (30) follows from randomization of X (Assumption 1 in the paper). From (30) it follows that (29) is correct if $Y(z, x) \amalg Z(x)$. This is, however, not the case for the DAG in Figure 1 of the paper (the path $Y(z, x) \leftarrow Q(x, z) \leftarrow - \rightarrow Z(x)$ in the corresponding twin network (Pearl, 2000) is open).

Known typos:

1. Page 2640, paragraph 2, line 3: '... than women who do not use HRT' should be '... than women who use HRT'.

2. Page 2647, two line above (A5): ≥ 1 should be ≤ 1.

5.2 Paper II

Physical activity (X) is known to be associated with both body mass index (Z) reduction, and a reduced risk for coronary heart disease (Y). In paper II we consider estimation of the direct (not mediated through Z) effect of X on Y. We focus on the PSDE. We assume that X can be considered randomized, conditional on a set of measured covariates, but that Z and Y may have unmeasured common causes. The scenario is illustrated by the DAG in Figure 7. For this scenario, the PSDE of X on Y is not identifiable. Gilbert et al (2003) have proposed a sensitivity analysis for the PSDE, which is based on a biased selection model, and could be used in this context. We propose a method of sensitivity analysis based on a pattern mixture model. Biased selection models and pattern mixture models are common tools for sensitivity analysis in the missing data literature (Molenberghs and Kenward, 2007). The difference between these two models lies in the parametrization. We argue that for sensitivity analysis of the PSDE, the pattern mixture model may be easier to interpret than the biased selection model. In Section 5 of the paper, we offer a detailed discussion of the pros and cons of the two models.

In our practical application, body mass index is dichotomized as 'obese' ($Z = 1$) or 'not obese' ($Z = 0$). Since we use this coarse version of body mass index, it may be problematic to interpret the PSDE as a 'direct' effect, as discussed in Section 4. To use a finer classification of body mass index would, however, introduce additional identification problems, as discussed in the paper.

5.3 Paper III

In paper III we consider estimation of the NDE of an exposure X, on an outcome Y. We assume that all three variables, X, Y, and the intermediate variable Z, are binary. We assume that X and $\{Z, Y\}$ have no common causes, but allow for Z and Y to have (unmeasured) common causes. The scenario is illustrated by the DAG in Figure 7. For this scenario, the NDE of

X on Y is not identifiable. We use a particular linear programming technique to derive bounds for the NDE. Our results extends those of Cai et al (2007) who derived bounds for the CDE under the same assumptions.

A technical remark
We claim in paper III (Section 2), but do not prove, that

$$\{Y(0,1), Y(0,1), Y(1,0), Y(1,1), Z(0), Z(1)\} \amalg X \tag{31}$$

implies that $\Pr\{Y(x) = y\} = \Pr(Y = y|X = x)$. Here follows a formal proof.

$$\begin{aligned}
\Pr\{Y(x) = y\} &= \sum_z \Pr\{Y(z,x) = y, Z(x) = z\} \\
&= \sum_z \Pr\{Y(z,x) = y, Z(x) = z|X = x\} \\
&= \sum_z \Pr(Y = y, Z = z|X = x) \\
&= \Pr(Y = y|X = x),
\end{aligned}$$

where the first equality follows from consistency and standard probability rules, the second from (31), the third from consistency, and the fourth from standard probability rules.

5.4 Paper IV

Several recent studies have reported that women who have used hormone replacement therapy (HRT), and developed breast cancer, tend to have a better prognosis than women with breast cancer who have not used HRT. One possible explanation is that tumors caused by HRT are more benign than tumors caused by other factors. Although it is relevant to quantify differences in prognostic factors across subtypes of breast cancer, it is not obvious how to do this correctly. This is because the tumors which occur among women who are treated with HRT are a mixture of HRT-induced and other tumors. In paper IV we combine the biological hypothesis that HRT only influences prognosis for women with HRT-induced tumors, with the framework of principal stratification, to discriminate between women with different subtypes of cancers.

In the literature on principal stratification, both the exposure, X, and the intermediate variable, Z, are often assumed to be binary (see Section

4). This means that the number of possible principal strata is $2 \times 2 = 4$. Our work is a novel contribution in that we manage to handle a continuous exposure, X. When X is continuous, there are 'infinitely' many principal strata, which makes the analysis more involved.

Remark The results in Appendix D of paper IV were derived by co-author Stijn Vansteelandt.

6 Discussion

Causal inference has been an intense research field in statistics for the past 10-20 years. Even so, most of the novel causal inference methods which have been developed have not yet found their way into the standard statistical toolbox. There may be several reasons for this. For a long time, the only way to phrase causal queries and assumptions was through counterfactual variables. Many people have difficulties in interpreting counterfactuals (see Section 2). More recently, DAGs have been proposed as a complement to counterfactuals, Their introduction has made the causal inference language more transparent and intuitively appealing. DAGs are likely to play a key role in a future, happy symbiosis between traditional statistics and the modern causal inference framework.

There is also a mathematical barrier to understanding the causal inference literature for many epidemiologists and statisticians. Many of the ground breaking papers in causal inference have been technical and mathematically involved. This may indicate that causal inference is not a trivial topic, but it may also reflect the personalities of the more productive researchers in the field. In any case, there is an obvious need for pedagogical overview papers and books, on a less technical level than the original papers, but without compromising too much with the mathematical foundation (see Hernan and Robins (2006) for an excellent example).

The slow acceptance of causal inference in statistics may also be due to the discrepancy between peoples' expectations on causal inference, and the natural limitations of what causal inference methods can achieve. In my experience, it is not uncommon that people think of causal inference as a 'magic toolbox' which allows those who are initiated to draw valid causal conclusions from any kind of data. On the contrary, applying a formal causal reasoning to the problem at hand often reveals the limitations of the study

design, and that the assumptions necessary for making causal inference are often unrealistically strong. Although such a message may be depressing, I view it as an argument *for* a broader use of causal inference methods, not against.

Acknowledgements

During the work with this PhD thesis I have had the pleasure to interact with many supporting people, who I would like to thank here.

My main supervisor, Keith Humphreys, my first co-supervisor, Juni Palmgren, and my second co-supervisor, Ola Hössjer. I have always felt that you have trusted me, and given me time and freedom to develop into an independent researcher.

I would like to thank Keith for always having time for discussions, for providing me with great ideas (in particular on paper IV, which I think we both are very pleased with), and for always giving me valuable feedback on my own ideas. You have also taught me to be very careful about the whole scientific process, and not letting anything be submitted unless it is really 'water proof'.

I would like to thank Juni for introducing me to the field of causal inference during my undergraduate studies. Without you I would probably have spent the past 5 years doing something much less interesting. I would also like to thank you for introducing me to an international scientific community; you gave me both the opportunity to study at Harvard and to give presentations at international statistical conferences.

I would like to thank Ola for providing me with valuable ideas and support during the work with a statistical genetics paper in the beginning of my PhD studies. The paper was published in *Annals of Human Genetics*, but was eventually not included in this thesis.

Stijn Vansteelandt, co-author on paper II and IV. Your help and guidance during this PhD work was indispensable. I have really learned a lot about causal inference methodology from our discussions.

Yudi Pawitan. I would like to thank you for always having time for my questions about statistical inference. I would also like to thank you for having written such an excellent book about likelihood inference. I learned a lot, both about statistics and about pedagogics, from reading it.

Marie Jansson. I would like to thank you for your wonderfully efficient and careful handling of all administrative details regarding my PhD position.

The whole biostat group at MEB. When I started my PhD studies I chosed to work at MEB because of the friendly environment, and the vivid scientific culture. I have never had any reason to regret this choice. In particular, I would like to thank my room mate Gudrun Jonasdottir, for being a good

friend and colleague, and for not complaining about my tuba, my trumpet, and my dirty gym-t-shirts, which were often spread across our room.

Finally, I would like to thank my girlfriend, Hanna, for always showing interest in my work, and for always supporting me. This work wouldn't have been half as fun to carry out without you!

References

Cai Z, Kuroki M, Pearl J, Tian J. (2007). Bounds on direct effects in the presence of confounded intermediate variables. *Biometrics* **64**(3), 695-701.

Frangakis CE, Rubin DB. (2002). Principal stratification in causal inference. *Biometrics* **58**, 21-29.

Gilbert PB, Bosch JB, Hudgens MG. (2003). Sensitivity analysis for the assessment of causal vaccine effects on viral load in HIV vaccine trials. *Biometrics* **59**, 531-541.

Greenland S. (1999). Identifiability. In *Encyclopedia of Biostatistics*. Armitage P, Colton T (eds). John Wiley and Sons Ltd.

Hernan MA, Robins JM. (2006). Instruments for causal inference - An epidemiologists dream? *Epidemiology* **17**(4), 360-372.

Joffe MM, Colditz GA. (1998). Restriction as a method for reducing bias in the estimation of direct effects. *Statistics in Medicine* **17**, 2233-2249.

Molenberghs G, and Kenward MG. (2007). *Missing data in clinical studies*. John Wiley and Sons Ltd, England.

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

Pearl J. Direct and indirect effects. (2001). In *Proceedings of the Seventeenth Conference on Uncertainty in Artificial Intelligence*, 411-420. San Francisco, CA: Morgan Kaufmann.

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

Robins JM. (1997). Marginal structural models. In *Proceedings of the American Statistical Association. Section on Bayesian Statistical Science*. 1-10.

Robins JM. (2003). Semantics of causal DAG models and the identification of direct and indirect effects. In *Highly Structured Stochastic Systems*, 70-81. Oxford: Oxford University Press.

Robins JM Rotnitzky A, Vansteelandt S. (2007). Discussion of Principal stratification designs to estimate input data missing due to death. *Biometrics* **63**, 650-654.

Rosenbaum PR, Rubin DB. (1983). The central role of propensity scores in observational studies for causal effects. *Biometrika* **70**(1), 41-55.

Rotnitzky A, Robins JM, Scharfstein DO. (1998). Semiparametric regression for repeated outcomes with nonignorable nonresponse. *Journal of the American Statistical Association* **93**(444), 1321-1340.

Rubin DB. (1974). Estimating causal effects of treatments in randomized and non-randomized studies. *Journal of Educational Psycology* **66**(5), 688-701.

Rubin DB. (2004). Direct and indirect causal effect via potential outcomes. *The Scandinavian Journal of Statistics* **31**, 161-170.

Rosenbaum PR, Rubin DB. (1983). The central role of propensity scores in observational studies for causal effects. *Biometrika* **70**(1), 41-55.

Verma T and Pearl J. (1988). Causal networks: Semantics and expressiveness. In *Proceedings of the 4th Workshop on Uncertainty in Artificial Intelligence*, 352- 359. Mountain View, CA.