# From clustered data to causal inference: new methodology motivated by the analysis of subfertility treatments

**Sylvie Goetgeluk**

**Promotoren:**
Prof. dr. Stijn Vansteelandt en Prof. dr. Els Goetghebeur

Proefschrift ingediend tot het behalen van de graad van
Doctor in de Wetenschappen: Wiskunde

Academiejaar 2007–2008

UNIVERSITEIT GENT

FACULTEIT WETENSCHAPPEN
Vakgroep Toegepaste Wiskunde en Informatica

# Dankwoord

In de eerste plaats wil ik mijn promotor Stijn Vansteelandt heel erg bedanken, niet alleen voor de intellectuele ondersteuning, maar ook voor het oneindig aantal uren die hij voor mij vrijgemaakt heeft en vooral voor zijn vertrouwen, zijn raad en zijn geduld. Hij is voor mij een grote steun geweest.

Ik wil mijn promotor Els Goetghebeur bedanken voor de intellectuele ondersteuning, voor de tijd die ze voor mij vrijmaakte en voor de vele nuttige commentaren.

Hartelijk dank aan de leden van de jury, voor hun tijd en interesse, en voor de nuttige commentaren.

Ik dank de mensen van het EFPTS, i.h.b. Catherine en Robert Derom, voor het gebruik van de tweelingen data. Dank aan dokter Lars Gram, professor Petra De Sutter en professor Christoph Lange voor gebruik van de andere datasets die in deze thesis aan bod komen.

Ook dank aan professor Marleen Temmerman, Ilse Delbaere en Hans Verstraelen van de dienst gyneacologie in het UZ Gent, aan professor Juni Palmgren en Arvid Shölander van het Karolinska Instituut en aan professor Krista Fischer en Heti Pisarev van de universiteit van Tartu voor de boeiende en vruchtbare samenwerking.

Ik dank Bieke, Ella, Ilse, Cynthia en Stijn, de beste collega's en vrienden die je je kan wensen, voor alle steun en moedgevende woorden, maar ook voor alle leuke middagen, avonden en weekendjes waar we ons telkens heel erg

amuseerden. Ik hoop dat onze vriendschap nog lang blijft duren. Speciale dank aan Ella en Bieke om mijn grafieken telkens in super tempo om te zetten van emf naar eps formaat. Om een of andere duistere reden wou mijn computer dit niet meer doen. Ook aan de andere (iets nieuwere) collega's van de onderzoeksgroep statistiek veel dank voor hun steun.

Cynthia, thanks for your support, for the nice time we had together in our office and for your rational view on many things which often leaded to good advice.

Ik dank ook mijn nieuwe collega's bij Itineris voor de steun, het begrip en de vele moedgevende woorden tijdens de laatste loodjes voor het verdedigen van dit werk.

Ik dank mijn ouders voor hun steun en geloof in mij en voor de vele kansen en mogelijkheden die ze mij geboden hebben.

Kjell en Stephanie, bedankt voor de reeds zoveel jaren durende hechte vriendschap, voor jullie begrip en hulp in moeilijke tijden, voor de ontspannende avondjes die me telkens weer oplaadden en voor de onvoorwaardelijke steun.

Laurence, bedankt voor je vriendschap, je interesse, je raad en hulp in moeilijke tijden.

Ik dank ook mijn zussen, mijn (schoon)familie en vrienden voor hun interesse en moedgevende woorden.

Lies en Joe, mijn allerliefste schatten, bedankt voor de reeds ontelbaar vele momenten van puur geluk die jullie mij met jullie stralende oogjes en schaterlachjes al geschonken hebben. Ze gaven me steeds weer de kracht om door te gaan.

En last but not least, Peter, mijn rots in de branding, ik kan jou niet bedanken met woorden. Ik kan alleen hopen dat mijn glimlach en liefde genoeg tonen hoe dankbaar ik je ben. Je bent nog steeds 'my kind', mijn zielsverwant. Het is dan ook aan jou en aan de 2 wondertjes die je mij (tot nu toe :) ) geschonken hebt, dat ik mijn thesis wil opdragen.

# Contents

**Preface** **1**

**1 Introduction to causal inference** **5**

1.1 Randomized trials . . . . . 8

1.2 Observational studies . . . . . 10

1.3 Causal directed acyclic graph (causal DAG) . . . . . 15

1.3.1 Graphical notation and terminology . . . . . 15

1.3.2 d-separation in practice . . . . . 20

1.4 Potential outcome model . . . . . 28

1.4.1 Individual causal effect . . . . . 29

1.4.2 Population causal effect . . . . . 30

**2 Analysis of twin data** **37**

2.1 Introduction . . . . . 37

2.2 Estimation of heritability . . . . . 38

2.2.1 Structural equation models . . . . . 41

2.2.2 Random effects models . . . . . 49

2.2.3 Heritability of birth weight in the East Flanders Prospective Twin Survey . . . . . 55

2.2.4 Heritability and confounding . . . . . 57

2.3 Estimation of causal exposure effects based on twin data . . 60

**3 Conditional generalized estimating equations for the analysis of clustered and longitudinal data 63**
3.1 Introduction . . . . . 64
3.2 Clinical effect of imipramine on depression . . . . . 66
3.3 Conditional generalized estimating equations . . . . . 68
3.4 Separating within- from between-cluster exposure effects . . 73
3.5 Data analysis . . . . . 76
3.6 Simulation study . . . . . 81
3.7 Discussion . . . . . 85

**Appendix 3.A1: Assumptions 86**

**Appendix 3.A2: proof of Theorem 3 88**

**Appendix 3.A3: proof of Theorem 4 92**

**Appendix 3.A4: comparison CGEE- versus NK-approach 95**

**4 Introduction to direct effect estimation 99**
4.1 Motivation for direct effects . . . . . 99
4.2 Definitions of direct effect . . . . . 101
4.3 The problem with inferring direct effects . . . . . 106
4.4 Case study: inferring the direct effect of ART on perinatal health other than through zygosity . . . . . 108
4.5 Structural equation models . . . . . 115
4.5.1 Inferring direct causal effects via structural equation models . . . . . 115
4.5.2 Simulation study . . . . . 116
4.6 Brief overview of literature about estimation of direct effects 124

**5 A general principle for the identification of direct causal genetic pathways in association studies 127**
5.1 Introduction . . . . . 128

5.2 Fallacies of intuitive regression adjustments . . . . . 131
5.3 A general principle to test for causal direct genetic effects . 135
5.4 Data analysis: An application to the Framingham Heart Study, the British Birth Cohort and the CAMP study . . . 139
5.5 Simulation Study . . . . . 141
5.6 Discussion . . . . . 150

**Appendix 5.A1: Distribution of the test statistic 151**

**6 Estimation of controlled direct effects 155**
6.1 Introduction . . . . . 156
6.2 Structural nested direct effect models . . . . . 159
6.2.1 Controlled direct effects . . . . . 159
6.2.2 Inverse Probability of Intermediate Weighted estimators . . . . . 159
6.2.3 Doubly-robust estimators . . . . . 164
6.2.4 Unweighted estimators and sequential G-estimators . 166
6.2.5 Stabilized doubly-robust estimators . . . . . 167
6.3 Simulation study . . . . . 169
6.4 Data analysis . . . . . 178
6.5 Discussion . . . . . 180

**Appendix 6.A1: Proof of Theorem 5 182**

**Final discussion and further plans 186**

**Bibliography 192**

**Nederlandse samenvatting 208**

# Preface

Detecting and quantifying cause-effect relations forms the basis for policy decisions and interventions in many fields of research. For this purpose, large amounts of data are gathered through experimental or observational studies and statistical techniques are used to analyze these data and to infer the effects of interest. However, standard statistical techniques are not directly aimed at inferring cause-effect relations. Instead, they are concerned with finding associations between measurements. Such associations may exist, even in the absence of a causal effect, and vice versa, when there exist prognostic variables for the outcome that also influence the exposure. Such variables are called confounders and they may complicate the estimation of a causal effect.

During the past three decades, important new insights have been obtained on how to infer causal effects, with seminal works by Don Rubin, James Robins and Judea Pearl. The main stimulus for many of these developments has been the introduction of potential outcome notation for causal effects (Rubin, 1978; Robins, 1986) and the development of causal diagrams for 'visualizing' causal effects (Pearl, 1995, 2000). The use of causal inference techniques has lead to important developments in statistics, such as on how to adjust for time-varying confounders in longitudinal studies. In Chapter 1, we give a brief introduction to causal inference with a main emphasis on causal diagrams and potential outcomes.

In this thesis, we will apply and develop causal inference methods for addressing substantive problems that were motivated through our collaborations with researchers at the Department of Obstetrics and Gynecology at the Ghent University Hospital, concerning twin data, infertility and peri-

natal outcomes.

In Chapter 2, we focus on twin data. Such data play an important role in medical research because they offer a unique source of information about the impact of genetic and environmental factors on human wellbeing. We first review structural equation models (SEM), which are commonly used for analyzing twin data and for estimating the impact of genetic factors (e.g. heritability) on phenotypes of interest. Estimation of heritability brings along the question whether heritability estimates may be confounded. To the best of our knowledge, this problem has only been tangentially addressed in the literature. At the end of Chapter 2, we therefore gain insight in this research question and offer a simple, novel way to obtain unconfounded heritability estimates.

Twin data are not only useful for estimation of genetic effects, they also have a rich structure for inferring causal effects because the comparability of twin children can be exploited to obtain effect estimates that are consistent in the presence of unmeasured confounders that are constant within twins, e.g. parental characteristics, environmental factors,... This principle is applicable for general clustered data. It is well known in the statistical literature, but is ignored by many frequently adopted methods (e.g. GEE with independence correlation structure). In Chapter 3, we develop a general methodology for clustered data with arbitrary correlation structure based on this principle. In particular, we develop semi-parametric efficient estimators for the parameters indexing marginal linear and loglinear models which include unmeasured confounders that are constant within clusters. On the basis of the resulting 'conditional generalized estimating equations', we study the validity of a simple adjustment procedure proposed by Neuhaus and Kalbfleisch (1998), which has been recommended for the analysis of twin data (Carlin et al., 2005).

Motivated by studies on the relative effect of subfertility treatments on perinatal health, from Chapter 4 onwards, we study the problem of separating an overall causal effect of an exposure on an outcome into an indirect effect through a given intermediate variable, and the remaining direct effect. In Chapter 4, we show that estimation of direct effects is a complex problem, although many of the complexities are ignored in stan-

dard practice. We argue that even the definition of a direct effect is very subtle. In addition, we consider structural equation models for inferring direct effects and motivate theoretically as well as through simulation that, in many realistic settings, such methods may yield severely biased direct effect estimates.

In Chapter 5, we develop direct effect estimators which are valid under weaker assumptions than traditional regression-based direct effect estimators. Simulation studies and application to a family-based genetic association study illustrate the dramatic improvements that can be realized by using this estimator for the direct effects instead of standard linear models. In Chapter 6, we show that these estimators are special cases of a general class of direct effect estimators which involve inverse probability weighting by a conditional distribution of the intermediate variable. We show that some of the estimators in our class can be very unstable when the intermediate variable is continuous. To obtain more stable and accurate inferences, we propose doubly robust estimators for direct effects. These estimators are asymptotically unbiased if either the model for a conditional density of the intermediate variable (i.e. the weights) is correctly specified or a model for a conditional expectation of the outcome. In addition, a number of doubly robust estimators are developed which are designed to behave relatively well in the presence of extreme weights. The different estimators are compared through extensive simulation studies and the analysis of perinatal data on singletons born after single or double embryo transfer.

Chapters 3, 5 and 6 were originally written as stand-alone articles. As a result, there exists a minor bit of overlap between these chapters. The notation is introduced per chapter and may therefore also differ throughout the complete thesis. Chapter 3 was published in Biometrics (Goetgeluk and Vansteelandt, 2008). Chapter 5 is under review for publication in the American Journal of Human Genetics (Vansteelandt, Goetgeluk et al., 2008) and thus, has a more applied focus. Chapter 6 is accepted for publication in the Journal of the Royal Statistical Society - Series B (Goetgeluk, Vansteelandt and Goetghebeur, 2008). The results have been presented at several international conferences.

# Chapter 1

# Introduction to causal inference

Detecting and quantifying cause-effect relations is the basis for policy decisions and interventions in many fields of research. In the medical and pharmaceutical sciences, for example, one is interested in the causal effect of a specific drug on a primary health outcome. In economics one wishes to investigate the effect of marketing campaigns on customer behaviour. In sociology one searches for causes of poverty and crime. Engineers question causes of failing of instruments in a satellite,... For this purpose, large amounts of data are gathered through experimental or observational studies and statistical techniques are used to analyse these data and to infer the effects of interest.

However, standard statistical techniques are not directly aimed at inferring cause-effect relations but are instead concerned with finding associations, correlations and dependency between an exposure and an outcome. Such associations and correlations may differ greatly from the causal effect of exposure on outcome. In fact, from data alone (i.e., in the absence of additional background knowledge about, for instance, the design of the study), one cannot infer whether an exposure affects an outcome, even when both are correlated. The following examples illustrate this.

(*Example 1*) Doll and Hill (1954) observed a high positive association between having tar-stained fingers and lung cancer mortality. Clearly having tar stains on one's fingers does not by itself cause lung cancer, but could be associated with lung cancer risk because smokers, who are at greater risk

of lung cancer, are also more likely to have tar-stained fingers.

(*Example 2* Dallal (2001)) During the Second World War, it was curiously noticed that bombers were less accurate when the weather was more clear. The reason was that when the weather was clear there was also more opposition from enemy fighter planes.

(*Example 3* Oberle et al. (2003)) People who develop asthma tend not to have cats because the presence of a cat aggravates their respiration. The negative association between having cats and the occurrence of asthma clearly does not indicate that cats have a protective effect on the risk of asthma.

It is clear that the associations found in these examples, do not reflect the causal effects of interest. We therefore call them spurious associations. Vice versa, causation also does not imply association.

(*Example 4* Delbaere et al. (2007a)) Maternal age at birth is negatively associated with birth weight among mothers with the same socio-economic status (the older the mother, the lower the birth weight of her child tends to be). Furthermore, mothers with a high socio-economic status tend to be more highly educated and therefore older when they get children. They also tend to be more healthy and thus more likely not to have low birth weight children. When these contrasting associations were of the same magnitude, no association would be found between maternal age and birth weight.

It does not often occur in practice that contrasting associations between an exposure and an outcome have exactly the same magnitude. It follows that when there is a causal effect of exposure on outcome, they will usually also be associated. The difficult part is then to rule out all spurious associations so that only the causal effect remains. How this can be done will be explained later.

Technically, that statistical association does not imply causation and vice versa, is because association between two events can be produced by several causal structures:

- When two events share a common cause, they will generally be associated, even if neither is the cause of the other. We then call this common cause a ‘confounder’ for the association between the two

events. In Example 1, tar-stained fingers and lung cancer mortality are both caused by smoking status, which is then a confounder for the effect of tar stains on lung cancer mortality. In the study by Delbaere et al. (2007a), socio-economic status affects both maternal age at birth and birth weight. Thus, it is a confounder for the association between them and may create a spurious association.

- When event 1 causes another event 3 which in turn causes event 2, event 1 and event 2 will generally be associated, even if event 1 does not directly cause event 2. In this case there is no direct but an indirect causal effect. Event 3 is then called a mediator or intermediate variable. In Example 2, clear weather has a causal effect on the occurrence of enemy fighter planes, which in turn affects the accuracy of the bombers. Thus, clear weather has an indirect effect on the accuracy of the bombers, mediated by the occurrence of enemy fighter planes.

- When event 2 is the cause of event 1, they will be associated, but event 1 is clearly not the cause of event 2. In Example 3, having asthma affects having cats and not the other way around. When cause and effect are interchanged, we talk about reverse causation.

- When two independent events have a common effect they will generally be associated within subgroups with the same occurrence of the common effect. This will be illustrated later in Example 5.

These scenario's/settings indicate an important difference between causation and association: that causation has a direction and association has not. Examples 1, 2 and 3 illustrate obvious mistakes or confusions and show how easy it is to make subtle errors when standard statistical analysis is used to prove causality in the absence of background knowledge. Such errors could have disastrous consequences if they form the basis of public policies and interventions. For example, Barret-Connor and Grady (1998) found that postmenopausal hormone therapy reduces the risk of coronary heart disease (CHD) in an observational study. The Women's Health Initiative, however, launched in 1991 and consisting of several randomized trials,

showed instead an increased risk of CHD (Manson et al. , 2003). Rossouw et al. (2007) recently showed that women taking hormone therapy in the randomized trial were older than those taking the therapy in the observational study. They found that women who initiated hormone therapy closer to menopause tended to have a reduced CHD risk compared with the increase in CHD risk among women more distant from menopause, although this trend did not meet their criterion for statistical significance. This example illustrates that large differences in conclusions can be obtained depending on the design of the study (e.g. observational or randomized) and on the population studied (e.g. older or younger patients upon initiation of hormone therapy). In the following sections, we elaborate on this difference between randomized and observational studies.

## 1.1 Randomized trials

The most persuasive evidence for establishing a causal relationship comes through experimental (randomized) studies in which investigators control the exposure. In randomized clinical trials for example, the exposure, such as a new medication, is allocated randomly to the study sample in such a way that the treated and untreated groups are otherwise equivalent, at least in expectation. It follows that, if this randomization process has been successful, differences between the treated and untreated groups must reflect the causal effect of treatment on outcome and not a spurious association.

Although randomized trials are simple in concept, proper execution in human populations is often quite challenging and complicated. Even if randomization is successful in assuring comparability of exposed and comparison groups, validity of results for causal inference is not assured. For example, it is well known that powerful placebo effects operate in humans. People might already feel better, just by thinking or believing they are treated. This effect can be eliminated by concealing treatment status from the study participants. In that case, both the people in the placebo group and in the treatment group are unaware of their assigned treatment and then, a fair comparison can be made. More subtle problems may arise

when the physicians or others administering the treatment and collecting the outcome data, know the treatment status. The resulting potential for bias has prompted the use of 'double-blind' designs in which neither the study participants nor the physician administering the treatment know the treatment status.

Problems may also arise in randomized trials when the study participants do not comply with their assigned treatment. If, for example, many patients in the experimental group do not take their treatment properly, the treatment may appear less effective. Patients may also drop out of the study before their outcome is ascertained. When these subjects are not exchangeable or comparable with subjects who remained through the end of the study, but form a selective subgroup, then the difference in outcome between the treated and untreated may not reflect the causal effect of exposure on outcome. For example, if a treatment is beneficial and healthier patients are more likely to leave the study, the treatment will appear less efficient from the observed data.

Finally, the usefulness of randomized studies is sometimes questioned because, certainly in clinical studies, the health conditions of the study subjects are more closely followed up than in real life, because compliance to the treatment is more attentively controlled and because researchers may choose to enroll the less severely affected subjects in the study. These conditions make the sample of patients in clinical studies possibly not entirely representative for the target population. An example for this is given by the comparison of intrapartum and neonatal single-dose nevirapine with zidovudine for prevention of mother-to-child transmission of HIV-1. The clinical trial conducted by Guay et al. (1999) in Kampala, Uganda found that nevirapine lowered the risk of HIV-1 transmission during the first 14-16 weeks of life by nearly 50% in a breastfeeding population. Thus, it was concluded that this simple and inexpensive regimen could decrease mother-to-child HIV-1 transmission in less-developed countries. However, a real life observational study conducted by Quaghebeur et al. (2004) in Kenya showed that the perinatal HIV-1 mother-to-child transmission rate at 14 weeks after use of nevirapine was 18.1%, similar to the 21.7% without the intervention. Such data raise the question whether further evaluation

of the simple nevirapine regimen in field conditions is necessary, and suggest that the conditions and results in clinical trials may not always be representative for conditions and results in real life.

Moreover, in general, randomized trials involving humans are also ethically limited in the range of questions to which they can be applied. Many of the major questions of public health, concerning for example effects of smoking behaviour on lung cancer[1] or effects of water and air pollution, cannot be addressed through randomized trials because it is not ethical to expose humans experimentally to smoke or other harmful substances. For such questions we are limited to and in need of passively observing the health of people naturally exposed, that is, to observational studies.

## 1.2 Observational studies

In observational studies, the investigator does not control the exposure of people in the study and does not intervene on the population under study, other than to take measurements. In these studies, conclusions are based on differences between exposed and unexposed groups of different individuals which, unlike in randomized trials, may lack comparability. For example, exposure status may be determined by where people live or work, what they eat, what social group they belong to or by many other factors that can also be associated with the outcome. Thus, these factors are common causes of the exposure and the outcome and they may confound the causal effect of exposure on outcome. For example, when estimating the effect of pollution on health, the place of living must be taken into account. Indeed, not only is the rate of pollution different between modern cities and countryside villages, people living in modern cities may also differ from people living in the countryside in terms of health for other reasons than differences in pollution. If we assume for example, that countryside people eat more

[1]However, note that randomized encouragement designs are possible in this context (and have been used (Mark and Robins, 1993, Permutt and Hebel, 1989)), whereby subjects are randomized over encouragement to quit smoking or not. Those who did not quit smoking after being encouraged to quit, are then viewed as non-compliers to the assigned treatment.

healthy and are more relaxed than city people, and that cities are more polluted, then the effect of pollution on health would be overestimated.

As already pointed out, the presence of common causes between exposure and outcome may render them associated even when the exposure does not affect the outcome (Example 1). Conversely, no association may be measured even when the exposure affects the outcome. This may happen when these common causes act to reduce the effect size. Observational studies are nonetheless capable and often the only option of providing evidence about the causal relationship between exposure and outcome. The example at the end of the previous section furthermore shows that observational studies may sometimes give more relevant results for the population than randomized trials since they are closer to real life situations.

### How to resolve the problem of spurious associations?

In Example 1 on lung cancer and tar-stained fingers, one can intuitively understand that the spurious association between tar-stained fingers and lung cancer, induced by smoking, can be removed by doing the estimation per group of people with the same smoking behaviour. If no other confounders than smoking are present, then within these groups, the risk of lung cancer will not differ between people with more or less tar stains on their fingers, showing that tar stains do not cause lung cancer. This strategy of comparing people with the same smoking behaviour is called 'adjusting' the analysis by smoking status. However, in Example 5 (see further) we will show that adjusting for a variable may also cause biased results in some situations. One of the most difficult problems in the causal analysis of observational studies is to find the covariates for which to adjust in order to obtain the causal effect of exposure on outcome and to determine whether it is valid to adjust for a covariate without biasing the result.

The elusive nature of adjustment was recognized as early as 1899, when Pearson and Yule discovered what is now called Simpson's paradox: that any statistical relationship between two variables may be reversed or negated by including additional factors in the analysis. One of the best

known real life examples of Simpson's paradox occurred when the University of California, Berkeley was sued for bias against women applying to graduate school. The admission figures for fall 1973 (see Table 1.1) showed that overall, male applicants were more likely than female applicants to be admitted, and the difference was so large that it was unlikely due to chance (Bickel et al. , 1975). However when examining the individual departments, it was found that no department was significantly biased against women; in fact, most departments suggested a small bias against men. The explanation turned out to be that women tended to apply to departments with low rates of admission, while men tended to apply to departments with high rates of admission. The conditions under which department-specific frequency data constitute a proper defense against charges of discrimination are formulated in Pearl (2000).

| Major | Men | | Women | |
|---|---|---|---|---|
| | # Applicants | % admitted | # Applicants | % admitted |
| A | 825 | 62% | 108 | **82%** |
| B | 560 | 63% | 25 | **68%** |
| C | 325 | **37%** | 593 | 34% |
| D | 417 | 33% | 375 | **35%** |
| E | 191 | **28%** | 393 | 24% |
| F | 272 | 6% | 341 | **7%** |
| overall | 8442 | **44%** | 4321 | 35% |

Table 1.1: *Admission percentages for different departments at the Unit of California, Berkeley by gender. Numbers in bold show were the highest percentage was found per department.*

Despite a century of analysis, Simpson's reversal phenomenon continues to 'trap the unwary' (Dawid, 1979, Pearl, 2000) and the main question whether an adjustment for a given covariate is appropriate in any given study, continues to be decided either informally, based on tradition or intuition, or either through (among others) one of the following possible strategies (Hernan et al. , 2002):

1) by building a regression model for the outcome given the exposure

and deciding to include a certain covariate when the p-value of the estimated effect of the covariate in the model is less than 5 or 10 %;

2) by building a regression model for the outcome given the exposure and deciding to include a certain covariate when the relative change in estimate of the exposure effect before versus after adjusting for that covariate is greater than 10 %;

3) by deciding whether the covariate is a confounder according to the following definition of a confounder: a confounder is associated with exposure, it is associated with outcome conditional on exposure and it is not in the causal pathway between exposure and outcome.

Although widely used, the above approaches fail, as I will show with the following example, because they try to uncover causation merely from statistical associations and, as already pointed out, associations are not causations.

(*Example 5* (Slone Epidemiology Unit Birth Defects Study) Hernan et al. (2002)) Suppose that we wish to estimate the effect of taking daily supplementation of folic acid during the first two months of pregnancy on neural tube defects of the infant. We also have information of stillbirth or therapeutic abortion and must decide whether the analysis should be adjusted for this covariate. The artificial data can be found in Table 1.2.

| | C=1 | | C=0 | |
|---|---|---|---|---|
| | D=1 | D=0 | D=1 | D=0 |
| E=1 | 19 | 8 | 24 | 231 |
| E=0 | 100 | 46 | 94 | 658 |

Table 1.2: *Slone Epidemiology unit Birth Defects Study, E=mother took/did not take daily supplementation with folic acid during first 2 months of pregnancy (1/0), C= stillbirth or therapeutic abortion (yes=1/no=0), D=infant with/without neural tube defects (1/0).*

Following the first approach above, we find that, in a logistic regression model for the expectation of neural tube defects given folic acid use and stillbirth/therapeutic abortion, the p-value of the association of stillbirth/

therapeutic abortion with neural tube defects is less than 0.001. Thus, following the above strategy 1, we report the adjusted association of folic acid use and neural tube defects and find an odds ratio of neural tube defects in women with versus without folic acid use equal to 0.8 (95% CI [0.53;1.20]). Following the second approach, we note that the unadjusted odds ratio is 0.65, which gives a relative change of 0.23 compared to the adjusted odds ratio. We thus also report the adjusted odds ratio with this approach. Following the third approach, we find the same result since stillbirth/therapeutic abortion is associated with folic acid use (odds ratio is 0.55 with 95% CI [0.34;0.86]), is associated with neural tube defects within the group of unexposed infants (mothers who did not take folic acid) with an odds ratio of 15.22 (95% CI [10.09;22.95]) and is not on the causal pathway between folic acid use and neural tube defects, since neural tube defects can not appear after stillbirth or therapeutic abortion. We will explain in the next section why the adjusted odds ratio does not represent the causal effect of taking folic acid on neural tube defects, but (if there are no (other) confounders) that the unadjusted odds ratio represents this effect. Nonetheless, most studies restrict the analysis to liveborns.

Today, the statistical community is becoming increasingly aware that the above ad-hoc approaches are inappropriate and that caution should be taken. It is important that the concepts of cause and effect, and methods to estimate causal effects receive appropriate attention in statistics courses because statistics is commonly used for inferring cause-effect relationships. Newspapers, radio, television, and the internet are filled with claims based on some form of statistical analysis: 'Calcium is good for strong bones', 'watching TV is a major cause of childhood and adolescent obesity', 'drinking coffee during pregnancy causes babies to have a low birth weight', 'eating certain yogurts helps improving digestion'... To know which claims are valid it is necessary to understand what it takes to establish causality in order to be an intelligent consumer of the 'truths' the world throws at us.

One of the reasons why causality has for many years been ignored in the statistical literature is that causality cannot be translated in the vocabulary of probability theory, which is the mathematical language of statistics. Two languages for causality have recently been proposed and

offer a way to infer causal effects: a graphical representation based on so-called causal directed acyclic graphs (Pearl, 1995) and Neyman-Rubin's potential response model (Rubin, 1974). A causal directed acyclic graph is relatively easy to use and to understand. Roughly, it is a graphical representation of all causal influences between all variables of interest in the study. The potential outcome model is more complicated to use and to understand, but it is more explicit in terms of how causal effects are defined and offers a useful mathematical formalism for inferring causal effects. We will introduce both formalisms in Section 1.3 and Section 1.4, respectively.

## 1.3 Causal directed acyclic graph (causal DAG)

### 1.3.1 Graphical notation and terminology

In this section, we discuss causal DAGs as introduced by Pearl (1995, 2000) (see also Greenland et al. (1999) and Robins (2001)). A graph consists of a set $V$ of nodes and a set $E$ of edges that connect pairs of nodes. The nodes correspond to variables and the edges denote a certain relationship that holds in pairs of variables.

A graph is directed when all edges are directed, i.e. when the edges are arrows starting in one node and pointing to another node. A path in a directed graph $G$ is a sequence of edges such that each edge starts with the node ending the preceding edge. In other words, a path is any unbroken, nonintersecting route traced out along the edges in a graph, which may go either along or against the direction of the arrows. If every edge in a path is an arrow that points from the first to the second node of the pair, we have a directed or causal path.

A directed graph is acyclic (i.e. a directed acyclic graph) when it contains no cycles. This means that it is not possible, starting from a certain node, to end up in the same node by following a directed path. The diagram thus excludes mutual causation or feedback processes where an arrow starts in a node $X$ and ends in a node $Y$ and another arrow starts in node $Y$ and ends in node $X$. [2]

[2]Note that this does not exclude longitudinal data where, for example, at different

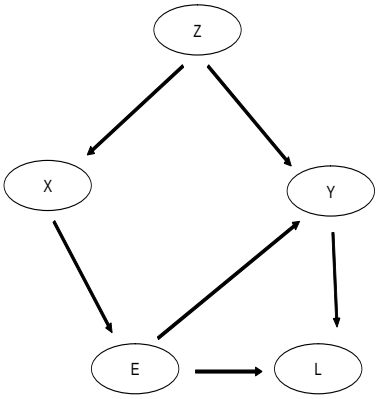


Figure 1.1: *Example of a DAG*

*Further terminology:*

- If there is an arrow from a node $X$ to another node $Y$ in the DAG, $X$ is called a **parent** of $Y$ and $Y$ is called a **child** of $X$.
- If there is a directed path from $X$ to $Y$, $X$ is called an **ancestor** of $Y$ and $Y$ is called a **descendant** of $X$.
- A path **collides** at a node $L$ if the path enters and exits $L$ through arrowheads, in which case $L$ is called a collider.
- Different types of paths between a node $X$ and a node $Y$:
  - A **directed path** from $X$ to $Y$, as explained before.
  - A **back-door path** from $X$ to $Y$: a path whose first edge is an arrow pointing to $X$ and whose last edge is an arrow pointing to $Y$.

time points $t = 1, 2, ...$, measurements on outcome $Y$ (i.e. $Y_1$, $Y_2$,...) and exposure $X$ (i.e. $X_1$, $X_2$,...) are taken. The DAG then contains all variables $Y_1$, $Y_2$,... and $X_1$, $X_2$,... and expresses that exposure at a given time may only affect outcome at later times.

– A **blocked path** between $X$ and $Y$: a path that has one or more colliders; otherwise it is **unblocked or open**. Thus, a back-door path and a directed path are open paths.

For example in Figure 1.1, $X$ is not a parent of $Y$ and $Y$ is not a child of $X$ since there is no arrow from $X$ to $Y$. However, there is a directed path from $X$ to $Y$ through $E$, so $X$ is an ancestor of $Y$ and $Y$ is a descendant of $X$. There is a back-door path from $X$ to $Y$ through $Z$ and a blocked path between $X$ and $Y$ since $L$ is a collider on the path from $X$ to $Y$ going through $E$ and $L$. Finally, there are 2 open paths between $X$ and $Y$; one through $Z$ (the back-door path) and one through $E$ (the directed path).
A DAG is causal (Pearl, 1995, 2000; Robins, 2001)

1) when every arrow in the DAG represents a stable and autonomous causal relation between the parent variable and the child variable, and

2) when the variables represented by nodes on the graph include the measured variables and additional unmeasured variables, such that if any two variables on the graph have a cause in common, that common cause is itself included as a variable on the graph, even if unmeasured.

In a causal DAG, a directed path represents a causal pathway, and an $X$-to-$Y$ arrow represents a direct effect of $X$ on $Y$ within the graph (an effect not mediated through any other variable in the graph). The absence of an arrow between two variables thus implies the assumption of no direct effect of any of the two variables on the other variable.

By representing the causal mechanisms and relations between variables, a causal DAG helps to get insight in the data and to express a priori background knowledge. Moreover, a simple tool, called d-separation, is available to determine on the basis of a causal DAG for which variables to adjust in an analysis in order to obtain the causal effect of an exposure on an outcome. To understand this tool and the possibility it offers to detect spurious associations between exposure and outcome, we first need to link the causal structure of the DAG to the statistical language of probability,

association and dependency. We will therefore introduce several definitions based on Pearl (2000), and then explain how d-separation works.

Suppose we have a set of $n$ variables $X_1, ..., X_n$ in a DAG. Then it follows by the chain rule that the probability of the joint event $(X_1, ..., X_n)$ can be written as a product of $n$ conditional probabilities:

$$P(X_1, X_2, ..., X_n) = P(X_n|X_{n-1}, ..., X_2, X_1)...P(X_2|X_1)P(X_1) \quad (1.1)$$

As the following theorem shows, the causal DAG imposes restrictions on this probability law.

**Theorem 1** (The Causal Markov Assumption (CMA) (Pearl, 2000))**.** *Any distribution generated by a causal DAG can be factorized as*

$$P(X_1, X_2, ..., X_n) = \prod_{i=1}^{n} P(X_i|PA_i) \quad (1.2)$$

*with $PA_i$ denoting the set of variables containing all direct causes of $X_i$ in the causal DAG.*

This assumption states that in a causal DAG, any variable that is not caused by a given variable $V$ will be independent of (i.e. unassociated with) $V$ conditional on the direct causes (i.e. the parents) of $V$. In other words, the CMA is the assumption that $V$ is independent of its nondescendants after adjusting for its parents.

**Definition 1.** *If a probability function $P$ admits the factorization of (1.2) relative to a DAG $G$, we say that $G$ and $P$ are compatible.*

The connection between the causal DAG and standard statistical dependence/association is made in the following theorem due to Verma and Pearl (1988), using the tool d-separation.

**Definition 2** (d-separation)**.** *Consider three disjoint sets of variables, $X$, $Y$ and $A$, which are represented as nodes in a causal directed acyclic graph. $X$ and $Y$ are said to be d-separated by a set of nodes $A$ if and only if each path between $X$ and $Y$ contains one of the following paths*

- *a directed path $i \to m \to j$ or a back-door path $i \leftarrow m \to j$ such that the middle node $m$ is in $A$;*

- *a blocked path $i \to m \leftarrow j$ such that the middle node $m$ is not in $A$ and such that no descendant of $m$ is in $A$.*

**Theorem 2** (Verma and Pearl, 1988)**.** *If a set of variables $X$ and another set of variables $Y$ are d-separated by a third set of variables $A$ in a DAG $G$, then $X$ is independent of $Y$ conditional on $A$ in every distribution $P$ compatible with $G$. Conversely, if $X$ and $Y$ are not d-separated by $A$ in a DAG $G$, then $X$ and $Y$ are dependent conditional on $A$ in at least one distribution $P$ compatible with $G$.*

The converse of Theorem 2 is in fact much stronger (Pearl, 2000): the absence of d-separation implies dependence in almost all distributions compatible with $G$. The reason is that a precise tuning of parameters is required to generate independency along a path, and such tuning is unlikely to occur in practice. This was already pointed out intuitively in Section 1.2 with the example of the study conducted by Delbaere et al. (2007a).

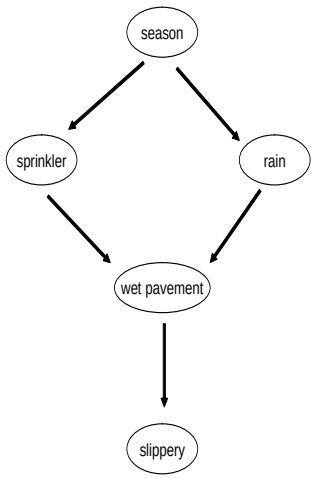


Figure 1.2: *Wet pavement example*

For example, examining the causal relationships among the five variables in the DAG in Figure 1.2, which is due to Pearl (2000), we see that rain and use of the sprinkler are d-separated by the season, meaning that after adjusting for season there remains no association between rain and use of the sprinkler. Intuitively, we can understand that once we know the season and assuming that the sprinklers are set in advance, according to the season, rain and use of the sprinkler are independent. However, since there is a closed path between rain and use of the sprinkler and wet pavement is the collider, finding that the pavement is wet or slippery (i.e. conditioning on the collider or its descendant) renders the rain and the use of the sprinkler dependent because refuting one of these explanations increases the probability of the other.

### 1.3.2 d-separation in practice

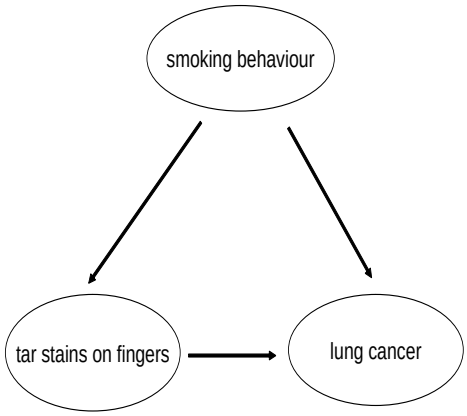


Figure 1.3: *Causal DAG corresponding to Example 1*

In practice, a causal DAG reflects and is based on expert knowledge of researchers in the field of study. Ideally researchers gathering data and statisticians analysing the data cooperate closely even before the data is

gathered and design a causal DAG to represent the mechanism they believe is generating the data as well as to represent design characteristics (e.g. ascertainment conditions). First, the effect of interest is displayed in the graph by drawing an arrow from exposure $X$ to outcome $Y$. Then, measurements that may affect both $X$ and $Y$ are added to the DAG. This is because all common causes of any two variables in the DAG should be included, since otherwise the DAG is not causal and cannot be used to infer the causal effect of exposure on outcome. We will see later, using d-separation, that precise knowledge of all the common causes may not be needed for inferring the effect of interest. When the analysis is restricted to a subset of the population (e.g. liveborns), the variable representing the different subsets should also be added to the DAG. Finally, proper attention should be paid to thinking about all possible relations between all variables in the DAG and the necessary arrows should be added.

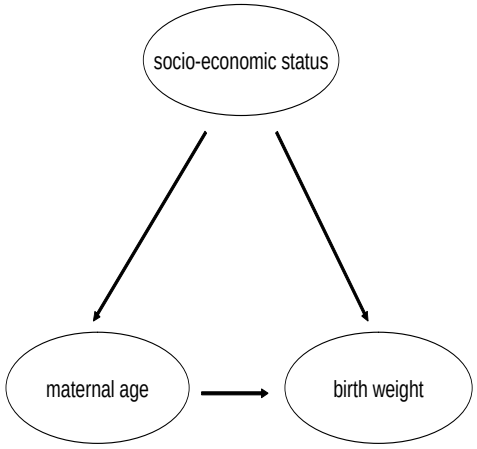


Figure 1.4: *Causal DAG corresponding to Example 4*

Assuming that there are no other variables affecting the variables in Examples 1, 4 and 5, the common cause problem of Example 1 is represented by the causal DAG in Figure 1.3 with an arrow from smoking behaviour to tar-stained fingers and to lung cancer, besides the arrow be-

tween smoking and lung cancer, which represents the effect of interest. The common cause problem of Example 4 is represented by the causal DAG in Figure 1.4 with an arrow from socio-economic status to birth weight and to maternal age, besides the arrow between maternal age and birth weight, which represents the effect of interest. The common effect in Example 5 is represented by the DAG in Figure 1.5 with an arrow from 'use of folic acid' to 'stillbirth/therapeutic abortion' and an arrow from 'neural tube defects' to 'stillbirth/therapeutic abortion'.

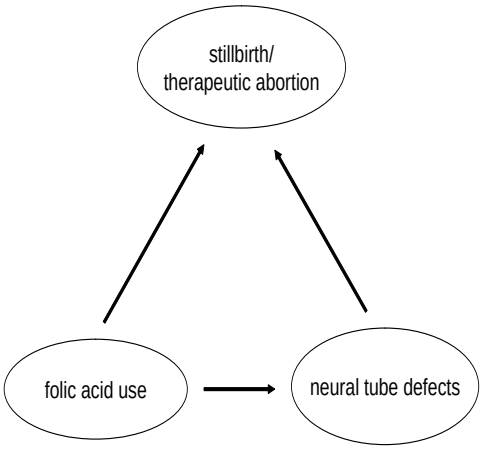


Figure 1.5: *Causal DAG corresponding to Example 5*

Once the DAG is considered complete (and thus, causal) it reveals which variables are needed to adjust for in the analysis and which variables should not be adjusted for in order to find the causal effect of the exposure on the outcome. It thus also reveals which variables should ideally be measured. This can be done by using the graphical tool called d-separation which will now be explained in a practical manner.

To know for which set of variables $A$ one should adjust to obtain the causal effect of exposure $X$ on outcome $Y$, we need to ascertain whether $X$ and $Y$ are d-separated conditional on $A$ after having removed the arrow of $X$ to $Y$ (i.e. the causal effect of interest). If they are d-separated conditional on

$A$ in the resulting DAG, it indicates that $X$ and $Y$ are no longer associated conditional on $A$, after having removed the causal effect of $X$ on $Y$, and thus that association is equivalent with causation. It thus suggests that it is necessary to adjust for $A$ to obtain the causal effect of $X$ on $Y$.

There are 3 practical rules to determine whether $X$ and $Y$ are d-separated or d-connected (Pearl, 2000)

1) $X$ and $Y$ are d-connected when there is an open path between them. Otherwise, they are d-separated.

2) $X$ and $Y$ are d-connected conditional on $A$ when there is a collider-free path between them that traverses no member of $A$. If no such path exists, then $X$ and $Y$ are d-separated by $A$. We also say then that every open path between $X$ and $Y$ is blocked by $A$.

3) If a collider is a member of the conditioning set $A$, or has a descendant in $A$, then it no longer blocks any path that traces this collider.

For example in Figure 1.1 we observe 3 paths between $X$ and $Y$; one open back-door path through $Z$, one open directed path through $E$ and one closed path through $E$ and $L$ in which $L$ is the collider. $X$ and $Y$ are not d-separated by $A$ if $A$ is an empty set. If $A$ contains the variables $Z$ (node on the open back-door path between $X$ and $Y$) and $E$ (node on the open directed path between $X$ and $Y$), then $X$ and $Y$ are d-separated by $A$. If $A$ includes the variable $L$, which is a collider in the closed path $X$-$E$-$L$-$Y$ between $X$ and $Y$, then $X$ and $Y$ are not d-separated by $A$. Thus, in this example, to obtain the direct effect of $X$ on $Y$ (which is no effect, according to the DAG) we should adjust for $Z$ and $E$ and not for $L$.

In Example 1 (Figure 1.3) we can now see that the exposure and outcome of interest are d-connected by another open path than the causal effect that we wish to obtain. To find the causal effects of interest we can block the open (spurious) path by conditioning on (i.e. adjusting for) smoking. In Example 5 (Figure 1.5) there is a closed path between folic acid use and neural tube defects and stillbirth/therapeutic abortion is the collider. Thus, adjusting for stillbirth/therapeutic abortion renders folic acid use and neural tube defects d-connected. Therefore, the adjusted odds

ratio calculated in this example does not represent the causal effect of folic acid use on neural tube defects. The unadjusted odds ratio, on the other hand, does give the causal effect (under the assumption that there are no confounders) because folic acid use and neural tube defects are d-separated.

### Inferring direct causal effects

In some cases, researchers wish to investigate the causal effect of exposure $X$ on outcome $Y$ which is not mediated through other variables $K$. For instance, in Example 2 the interest lies in estimating the causal effect of weather on the accuracy of bombers that is not mediated through the occurrence of enemy fighter plains. We call this the direct causal effect (or direct effect). We will discuss this type of effect in more detail in Chapter 4. A formal definition of a direct effect will be given in Chapter 5 and 6. The effect of $X$ on $Y$ mediated through other variables $K$ is called an indirect effect and is represented by a directed path from $X$ to $Y$ which contains at least 2 edges. The variables $K$ intermediate on such paths, are called intermediate variables. When drawing the DAG in that case, the intermediate variables are also added to the DAG together with arrows from $X$ to $K$ and from $K$ to $Y$. Then, for each pair of variables, measurements which affect both variables are added to the DAG since otherwise, the DAG is not causal. Finally, necessary arrows representing all possible relations between these variables are added. After removing the arrow from $X$ to $Y$ representing the direct effect, the 3 practical rules in the previous section can again be used to assess whether $X$ and $Y$ are d-separated conditional on a set of variables $A$. The additional 'spurious' associations along the open directed paths between the exposure and the outcome will be blocked then, meaning that one node along such directed path will be included in the set of variables $A$ and thus, this variable will be adjusted for in the analysis.

In Example 2, the indirect causal effect of 'clear weather' on 'accuracy of the bombers' is represented by the causal DAG in Figure 1.6 with a directed path starting with an arrow from 'clear weather' to 'enemy fighters', followed by an arrow to 'accuracy of the bombers'. To find the direct

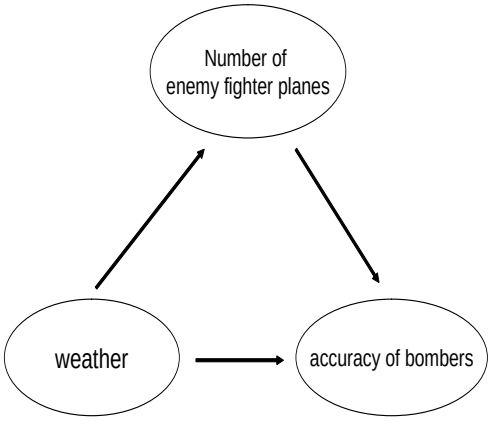


Figure 1.6: *Causal DAG corresponding to Example 2*

causal effect of 'clear weather' on 'accuracy of the bombers', d-separation indicates that we must block the open directed path through 'enemy fighters' by conditioning on (i.e. adjusting for) it.

### Problems that may occur

An unfortunate but realistic problem is that in many studies certain variables are unmeasured which, according to the DAG and after applying the d-separation rules, should be adjusted for in order to obtain the causal effect of an exposure on an outcome. This may sometimes be avoided if researchers first set up the causal DAG and examine which variables are needed for the analysis before the start of the study. Unfortunately, some of these variables are often difficult or expensive to measure, or some confounders may be unknown.

In certain specific settings this problem can (partially) be overcome by using special statistical methods which allow for the presence of (some) unmeasured confounders. The instrumental variables (IV) methodology (Heckman, 1979, Robins and Tsiatis, 1991; Robins, 1994; Goetghebeur and

Lapp, 1997; Vansteelandt and Goetghebeur, 2003) is one example. Here, prognostic variables (called instrumental variables) that have an unconfounded association with the outcome of interest and can only affect this outcome indirectly by modifying the target exposure, are used to overcome the necessity of knowing all confounders. Sometimes, these instrumental variables are the result of the design. In a double blind randomized study, for example, randomization is an instrumental variable because it affects the exposure, but does not affect the outcome (other than through the exposure). When estimating the effect of smoking on lung cancer, the price of cigarettes can function as an instrumental variable since it influences smoking behaviour, but has no direct effect on (getting) lung cancer. In Chapter 3 we will develop another class of methods which (partly) overcomes the problem of unmeasured confounders by making within-cluster comparisons of clustered data, e.g. twin data, family data, multicenter studies,....

When estimating the direct causal effect of an exposure on an outcome (which is not mediated through a third measurement), it frequently occurs that d-separation requires adjusting for a certain covariate to block a spurious association, but that by doing so a new spurious association is created. For instance, in Example 4, to estimate the direct causal effect of maternal age on birth weight that is not mediated through zygosity (monozygotic or dizygotic), we need to add zygosity to the DAG in Figure 1.4. Because type of conception (spontaneous or through artificial reproductive techniques) is a common cause of zygosity and birth weight, it must additionally be included. We thus draw the DAG in Figure 1.7 and find that it is necessary to block the open path through socio-economic status and the open directed path through zygosity. However, doing the latter opens the closed path between maternal age and birth weight through zygosity and type of conception since zygosity is a collider on that path. When type of conception is a measured variable in the study, this forms no problems as additionally adjusting for it besides zygosity resolves the problem. However, if type of conception were not measured, the results could be biased. It is frequently so that intermediate variables share common, possibly unmeasured, causes with the outcome. In the family-based association study conducted by Lyon et al. (2004) for example, the genetic association of certain SNPs in

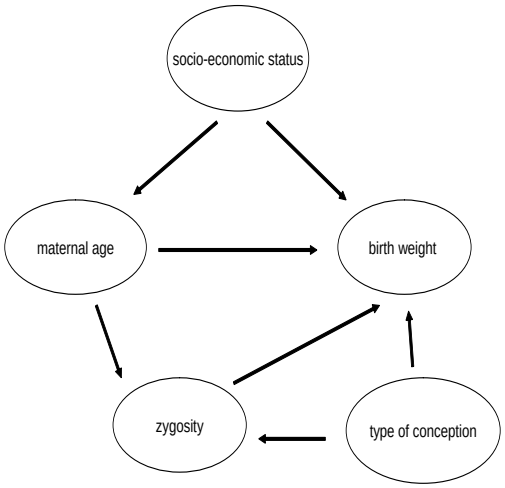


Figure 1.7: *Causal DAG for estimating direct causal effect of maternal age on birth weight*

the genome with asthma is investigated. It is found that certain SNPs that are associated with asthma are also associated with body mass index, which in turn affects asthma. To obtain the direct causal effect of each SNP on asthma it is thus necessary to adjust for body mass index. Doing so may be prohibiting because body mass index and asthma share common causes like gender, age, doing sports,..., some of which may not be measured.

In the study conducted by De Sutter et al. (2006), the effect of single embryo transfer (SET) versus double embryo transfer (DET) on perinatal outcomes is examined. SET/DET also affects gestational age which in turn affects birth weight. To obtain the direct causal effect of SET/DET on perinatal outcomes that is not mediated through gestational age, it is thus necessary to adjust for gestational age. Again, there are both measured and unmeasured common causes of gestational age and birth weight (e.g. vaginal blood loss during pregnancy, early contractions). Not being able to adjust the analysis for all common causes, may render the results biased. This example is explained in more detail in Chapter 6. Solutions to this problem are developed in Chapters 5 and 6.

Finally, note that a causal diagram merely gives insight for which variables it is necessary to adjust. Whether such adjustment is successful depends on whether the adjustment is correctly done. For example, when a confounder has a quadratic association with the outcome but enters linearly in the considered regression model, the estimated effect may be biased.

## 1.4 Potential outcome model

Although graphs are a very simple, transparent and non-parametric tool to discover confounders and spurious associations between an exposure and an outcome, a more quantitative or mathematical approach is often desired. For example, when developing statistical models with parameters that carry a causal interpretation, we need to define precisely what a causal effect is and ideally have notation for it. This is offered by the potential outcome or counterfactual model.

To infer the causal effect of exposure on outcome, we need to establish whether the outcome would have been different had the exposure been different, all other conditions being the same. For example, suppose that a child who lives near a chemical factory contracts a rare cancer. Suppose that we seek to establish whether or not a chemical spill adjacent to the child's property was the cause of his particular cancer. By saying that the chemical spill caused the disease on this individual level, we mean that the cancer would not have occurred had, contrary to fact, the spill not happened. When we investigate on a population level, the causal effect of having tar-stained fingers on lung cancer, we need to establish for people with tar stains, whether their lung cancer risk would have been different had they worn gloves, for instance, to protect their fingers from tar stains, all other things (like smoking behaviour) staying the same. It is clear from these examples that the ideal setting to inferring the causal effects is to examine the same people under different exposures, all other things staying the same.

The potential outcome model or counterfactual model formalizes this idea. In the following sections, we introduce this model, thereby making a distinction between individual causal effects and population causal effects

as in Hernan (2004) and Greenland and Brumback (2002).

### 1.4.1 Individual causal effect

Suppose that a person suffering from high blood pressure receives an experimental medication (i.e. the exposure $X$) and that we wish to establish its causal effect on blood pressure (i.e. the outcome $Y$). We define $Y_x$ as the outcome that person would have had if he/she received exposure $X = x$. Suppose that $X = 1$ corresponds to receiving the experimental medication and $X = 0$ correspond to receiving a placebo. Then $Y_1$ would have been the outcome for that person if he/she had received the experimental medication and $Y_0$ would have been his/her outcome if he/she had received the placebo. Since the person received the experimental medication, we can reasonably assume that $Y_1$ for that person equals his/her observed outcome $Y$. This assumption is commonly referred to as the consistency assumption. $Y_0$ is unobserved for this person. The variables $Y_1$ and $Y_0$ are called potential outcomes because one of them (namely $Y_0$) describes the subject's outcome value that would have been observed under a potential exposure value (i.e. $X = 0$) which differs from the actually observed exposure level. Because one of these outcomes would have been observed in situations that did not actually happen (that is, in counter to the fact situations), they are also called counterfactual outcomes (Rubin, 1978; Robins, 1986; Hernan, 2004).

The individual causal effect for a given person of treatment ($X = 1$) versus no treatment ($X = 0$) can now be defined as the difference

$$Y_1 - Y_0$$

between potential outcomes, corresponding to these different exposure levels. If we would observe both potential outcomes $Y_1$ and $Y_0$, this effect would be easy to calculate and causal inference would be simple. Unfortunately, in reality, we observe either $Y_1$ or $Y_0$ but not both. This is even so in cross-over studies where subjects receive both treatments but at different times so that not all conditions are identically the same. Studying monozygotic twins who have a different exposure at the same time only partly

solves this problem. In view of genetics, these twins are identical, but they do not experience the exact same environmental factors, which again involves possible different conditions. It follows that individual causal effects are never identified without very restrictive assumptions. In epidemiologic and scientific contexts, the interest is usually not so much in individual causal effects but the goal is to establish whether, in a population, certain exposures result in a change in the frequency or expectation of an outcome. The corresponding population causal effects can be identified under much weaker assumptions.

### 1.4.2 Population causal effect

Throughout this section, we will adapt an example from Hernan (2004) for illustration. Suppose we have a dichotomous exposure $X$ (for example $X = 1$ if a baby was born after a double embryo transfer (we will call this a DET baby) and $X = 0$ if a baby was born after a single embryo transfer (we will call this a SET baby) and a dichotomous outcome $Y$ (for example $Y = 1$ if a baby was born preterm and $Y = 0$ if a baby was born aterm). We define the probability $\mathrm{P}(Y_x = 1)$ as the proportion of babies that would have been preterm had they received 'treatment' (exposure) $x$. In this example, the exposure then has a population causal effect (or a causal effect for short) if $\mathrm{P}(Y_1 = 1) \neq \mathrm{P}(Y_0 = 1)$.

Suppose for instance that the population is comprised by the subjects in Table 1.3 and that all potential outcomes are observed. Then $\mathrm{P}(Y_1 = 1) = 10/20 = 0.5$ and $\mathrm{P}(Y_0 = 1) = 10/20 = 0.5$. That is, 50% of the babies would have been preterm had they all been DET and 50% would have been preterm had they all been SET. In that case, the exposure has no causal effect on the outcome at the population level. If a risk difference is chosen to report the causal effect, it equals $\mathrm{P}(Y_1 = 1) - \mathrm{P}(Y_0 = 1) = 0.5 - 0.5 = 0$.

In reality however, only one of both potential outcomes is observed. Specifically, the observed data for the example are those in Table 1.4. The question is then how to infer the causal effect of interest, despite the missing data on either $Y_0$ or $Y_1$.

When measuring the association between exposure $X$ and outcome $Y$,

| Subject | $Y_0$ | $Y_1$ |
|---|---|---|
| Rheia | 0 | 1 |
| Kronos | 1 | 0 |
| Demeter | 0 | 0 |
| Hades | 0 | 0 |
| Hestia | 0 | 0 |
| Poseidon | 1 | 0 |
| Hera | 0 | 0 |
| Zeus | 0 | 1 |
| Artemis | 1 | 1 |
| Apollo | 1 | 0 |
| Circe | 0 | 1 |
| Ares | 1 | 1 |
| Athene | 1 | 1 |
| Eros | 0 | 1 |
| Aphrodite | 0 | 1 |
| Prometheus | 0 | 1 |
| Selene | 1 | 1 |
| Hermes | 1 | 0 |
| Eos | 1 | 0 |
| Helios | 1 | 0 |

Table 1.3: *Counterfactual outcomes of subjects in a study with dichotomous exposure $X$ and outcome $Y$ (Hernan, 2004)*

we calculate the proportion of babies that were preterm in the subgroup that underwent treatment $X = 1$ (i.e. double embryo transfer) and compare this with the proportion that was preterm in the subgroup that underwent treatment $X = 0$ (i.e. single embryo transfer). Formally, one thus examines whether $\mathrm{P}(Y = 1|X = 1)$ equals $\mathrm{P}(Y = 1|X = 0)$. Looking at Table 1.4 we find that $\mathrm{P}(Y = 1|X = 1) = 7/13$ differs from $\mathrm{P}(Y = 1|X = 0) = 3/7$. It follows that there is an association between exposure $X$ and outcome $Y$. If a risk difference is chosen to report the association, it equals $\mathrm{P}(Y = 1|X = 1) - \mathrm{P}(Y = 1|X = 0) = 7/13 - 3/7 = 0.11$ which differs from the causal effect of zero.

| Subject | $X$ | $Y$ | $Y_0$ | $Y_1$ |
|---|---|---|---|---|
| Rheia | 0 | 0 | 0 | ? |
| Kronos | 0 | 1 | 1 | ? |
| Demeter | 0 | 0 | 0 | ? |
| Hades | 0 | 0 | 0 | ? |
| Hestia | 1 | 0 | ? | 0 |
| Poseidon | 1 | 0 | ? | 0 |
| Hera | 1 | 0 | ? | 0 |
| Zeus | 1 | 1 | ? | 1 |
| Artemis | 0 | 1 | 1 | ? |
| Apollo | 0 | 1 | 1 | ? |
| Circe | 0 | 0 | 0 | ? |
| Ares | 1 | 1 | ? | 1 |
| Athene | 1 | 1 | ? | 1 |
| Eros | 1 | 1 | ? | 1 |
| Aphrodite | 1 | 1 | ? | 1 |
| Prometheus | 1 | 1 | ? | 1 |
| Selene | 1 | 1 | ? | 1 |
| Hermes | 1 | 0 | ? | 0 |
| Eos | 1 | 0 | ? | 0 |
| Helios | 1 | 0 | ? | 0 |

Table 1.4: *Data and observed counterfactual outcomes from a study with dichotomous exposure $X$ and outcome $Y$ (Hernan, 2004)*

Note that the risk $\mathrm{P}(Y = 1|X = 1)$ is computed using the subset of subjects of the population that actually received the exposure $X = 1$ (that is, it is a conditional probability), whereas the risk $\mathrm{P}(Y_1 = 1)$ is computed using all subjects of the population had they received the (possibly) counterfactual exposure $X = 1$ (that is, it is an unconditional or marginal probability). Therefore, association is defined by a comparison of risks in two disjoint subsets of the population determined by the subjects' actual exposure value, whereas causation is defined by a comparison of risks in the same subset (for example, the entire population) under two potential exposure values. This different definition accounts for the well known adage

'association is not causation' (Hernan, 2004).

It follows from the above reasoning that when $P(Y_1 = 1)$ equals $P(Y = 1|X = 1)$, the causal effect of an exposure on an outcome can be obtained as the association between them. Causal assumptions are therefore naturally expressed in terms of independence assumptions between exposure and counterfactual outcomes as we exemplify in the next paragraph.

**Randomized trials**

In randomized trials, the treated and untreated groups are comparable (under the assumption of perfect compliance). This implies that if subjects were randomly assigned to group A and B, the proportion of subjects with outcome $Y = 1$ among the exposed will be the same whether these exposed are the subjects in group A or the subjects in group B. Thus, which particular group got the exposure is irrelevant for the value of $P(Y = 1|X = 1)$. Formally, we say that both groups are exchangeable. That is, the chance of $Y = 1$ in group A would have been the same as the chance of $Y = 1$ in group B had subjects in group A received the exposure given to those in group B. We may thus conclude that the risk under the potential exposure value $x$ among the exposed, $P(Y_x|X = 1)$, equals the risk under the potential exposure value $x$ among the unexposed, $P(Y_x|X = 0)$. Since these conditional chances are equal in all subsets defined by exposure status in the population, they must equal the marginal risk under exposure value $x$ in the whole population, i.e. $P(Y_x|X = 1) = P(Y_x|X = 0) = P(Y_x)$ and thus, $Y_x$ is independent of $X$ (notation $Y_x \coprod X$) for all values of $X$. Since we assume that $Y_x = Y$ for subjects actually receiving exposure $X = x$, we find that, under ideal randomized experiments (i.e. double blind, no loss to follow up, perfect compliance to treatment,...), $P(Y|X = 1) = P(Y_1)$ and $P(Y|X = 0) = P(Y_0)$. Thus, we can calculate the causal effect of the exposure on the outcome (e.g. the risk difference $P(Y_1) - P(Y_0)$) by calculating the association between them (i.e. $P(Y|X = 1) - P(Y|X = 0)$).

### Observational studies

In observational studies, the treated and untreated groups are generally not comparable and thus, not exchangeable. We thus cannot estimate causal effects merely by calculating associations without making additional assumptions. Subjects who are treated may differ from the untreated subjects in other things than just the treatment status. In that case, we expect their potential respons $Y_x$ to the same treatment $x$ to be different. This lack of exchangeability in observational studies is exhibited in $Y_x$ not being independent of $X$. The causal effect of exposure on outcome is then not the same as the association between them, i.e. $\mathrm{P}(Y|X = 1) \neq \mathrm{P}(Y_1)$ and $\mathrm{P}(Y|X = 0) \neq \mathrm{P}(Y_0)$. The association between exposure and outcome is then confounded or spurious.

Remember that d-separation can be used to verify independence assumptions. In particular, it can be used to verify whether the structure of a causal diagram is compatible with the assumption that $Y_x \coprod X$. However, since the potential outcome $Y_x$ is not shown on a causal DAG we need a way to make it explicit. Since $Y_x$ represents the outcome under a certain (fixed) value of $X$, the exposure $X$ itself cannot affect this potential outcome. This suggests that we can replace the outcome $Y$ by the potential outcome $Y_x$ in the DAG, provided that we remove the arrow from $X$ to $Y$. Now, we can use d-separation to find whether $Y_x \coprod X$. If $Y_x$ is not independent of $X$ then d-separation can be used to find for which covariates $A$ one needs to adjust so $Y_x \coprod X$ conditional on these covariates $A$, i.e. $Y_x \coprod X|A$. If we can find a set of covariates for which $Y_x \coprod X|A$, then, following the same reasoning as in the previous section, we find that $\mathrm{P}(Y|X = 1, A) = \mathrm{P}(Y_1|A)$ and $\mathrm{P}(Y|X = 0, A) = \mathrm{P}(Y_0|A)$. Thus, again, the causal effect of exposure $X$ on outcome $Y$ can be obtained by calculating the conditional association between them.

We will illustrate this with an example. Suppose we wish to calculate the causal effect of maternal age on birth weight (Example 4) represented by the DAG in Figure 1.4. Note that there is a back-door path between maternal age and birth weight through socio-economic status. This suggests that socio-economic status is a confounder for the association between

maternal age and birth weight. We can now replace outcome birth weight $Y$ by the corresponding potential outcome $Y_x$ in which $X$ represents the age of the mother, provided that we remove the arrow from maternal age to potential outcome $Y$. We then get the DAG in Figure 1.8. Using d-separation rules we now conclude that, in order for $Y_x$ to be d-separated, and thus, independent, of $X$, we need to adjust for socio-economic status and for diabetes.

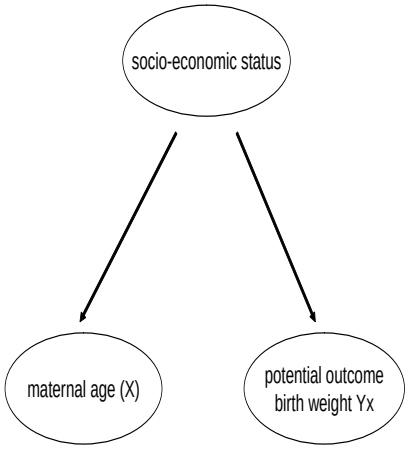


Figure 1.8: *Potential outcome in a causal DAG*

In conclusion, the population causal effect is obtainable under the following conditions:

- A causal DAG is created and all assumptions of such a DAG should be met.
- A set of variables $A$ should be found so that $Y_x \coprod X|A$.
- These variables $A$ should be accurately measured and correctly adjusted for.

A final remark on both the causal DAG methodology and the potential outcome model is that they rely on the assumption of no unmeasured con-

founders for the effect of exposure on the outcome (for estimating the total effect) and on the additional assumption of no unmeasured confounders for the intermediate variable and the outcome (for estimating the direct effect). These assumptions cannot be tested based on the observed data, which is a drawback of the methods. This might be the reason why many scientists are a bit sceptical towards these methods. However, causal DAGs allow to easily detect confounding problems, especially in complex settings like, for example, estimating direct causal effects where there are confounders of the intermediate variable and the outcome that are affected by the exposure. Causal DAGs also help to better understand complex ascertainment schemes (Robins et al., 2001).

In the following chapters, we will apply the ideas of causal DAGs and potential outcomes to address substantive problems that were motivated through collaborations with gynecologists at the Ghent University Hospital who are dealing with twin data, infertility and perinatal outcomes. In Chapter 2, we will examine the analysis of twin data, which offer a unique source of information about genetic and environmental factors, and we will address confounding problems in this setting. In Chapter 3, we will develop a methodology for analysing general clustered data which protects estimates against certain confounders. From Chapter 4 onwards, we will address the difficulties of estimating direct causal effects, starting with an introductory chapter and followed by two chapters in which solutions are offered.

# Chapter 2

# Analysis of twin data

## 2.1 Introduction

The development of causal methods in the statistical literature has so far mostly focused on data obtained from independent observation units, with some exceptions (Loeys, Vansteelandt and Goetghebeur, 2001; Vansteelandt, 2007; Albert, 2002). In practice, however, many other forms of dependency frequently arise. In twin and family studies, for example, measurements obtained on individuals from the same twin pair of family tend to be more alike. Similarly, studies that measure the effect of exposures on eyes, kidneys, teeth,... also deliver correlated data structures. In experimental studies, correlated data typically arise in cluster-randomized studies and multicenter studies. For instance, Sommer et al. (1986) analyse data on children who were (cluster-)randomized per village in rural Indonesia to either vitamin A or placebo. Here, children within the same treatment group are more alike by the fact that they are from the same village. Hirano et al. (2000) study the effect of flu vaccination and must deal with the fact that patients visiting the same doctor may be more alike. Clustered data are also very common outside the biomedical context: educational research is typically performed over different schools or classes, marketing studies try out different strategies in different shops or offices, machines are used under different circumstances in different periods of time,...

Ignoring correlation within clusters of data typically causes biased con-

fidence intervals for parameters of interest. The chance that these intervals cover the target population parameter is then not guaranteed to equal the level that was a priori specified. Depending on the data-generating mechanism and the parameter of interest, the calculated intervals may be too tight, in which case the analysis could be misleading, or too wide, in which case the analysis may be insufficiently informative. The former is typically true when assessing the effects of between-cluster exposures (i.e. exposures that do not vary within the cluster); the latter is more typical when estimating the effects of within-cluster exposures (i.e. exposures that do vary within the cluster). In both cases the analysis is not correctly reflecting the available information. With the goal of drawing valid inferences, a large collection of statistical methods is now available to accommodate the correlated nature of data in diverse application settings (Diggle, Liang and Zeger, 1994, Laird and Ware, 1985, Verbeke and Molenberghs, 1997, Verbeke and Molenberghs, 2000).

While correlated data structures call for a more complex analysis, they are frequently not only tolerated, but often also exploited to good use by statisticians/scientists. Indeed, by comparing differently exposed subjects from the same cluster, one may obtain more valid causal inferences by the fact that such subjects are more alike to begin with. More formally, such within-cluster comparisons allow to correct for unmeasured between-cluster confounders (i.e. unmeasured confounders that have a constant value within the cluster). In Chapter 3, we will expand on this and develop a general methodology for making within-cluster comparisons when analyzing clustered data. In this chapter, we focus specifically on the analysis of twin data.

## 2.2 Estimation of heritability

Twin data play an important role in medical research because they offer a unique source of information about the impact of genetic and environmental factors on human wellbeing. The impact of these factors is hard to detect on the basis of individual observations on independent units. The reason why twin data is so informative about the distinct roles of ge-

netic and environmental factors, lies in the fact that there are two types of twins: monozygotic and dizygotic twins. Monozygotic (MZ) twins, also called identical twins, are the result of division of a single fertilized ovum (zygote) at an early stage of development. Two individuals of identical genetic structure are therefore produced. Dizygotic (DZ) twins or fraternal twins are derived from two distinct fertilized ova. Like full siblings, DZ twins have, on average, half their genes in common.

The argument that data on MZ and DZ twins can be used to separate genetic from environmental influences is based on the following assumption:

**Assumption 1** (Equal environment)**.** *MZ and DZ twins do not differ in total environmental variance, or in the proportion of environmental variance that is common to members of the same twin-pair.*

Under this equal environment assumption, any excess similarity between MZ twins over that between DZ twins must be due to the greater proportion of genes shared by MZ twins than by DZ twins. In the following section, we describe two methods that exploit this idea to estimate the impact of genetic and environmental factors on human wellbeing.

We first note, however, that caution should be taken because the equal environment assumption may well be violated in realistic settings. Sham (1998) describes the following example. Suppose that a student of human heredity should hail from another planet and that he should be required to use the twin method to find out whether or not people's clothes were a direct consequence of heredity. He would find that identical twins were often dressed alike, often down to quite small details, and that this was uncommon with fraternal twins. He would confidently conclude that the choice of clothes was almost an exclusively hereditary trait. However, this conclusion cannot be trusted since the assumption of equal environment, on which the analysis is based, is not valid in this situation. This example illustrates how the environment can exaggerate the genetic component. There are various other situations in which it is realistic to believe that MZ twins (who are often more close than DZ twins) share more common environment, resulting in a violation of the assumption. Suppose, for example, that lung cancer is no hereditary trait. Since MZ twins likely have a more

similar smoking behavior than DZ twins (e.g. as a result of sharing more friends), the occurrence of lung cancer will also be more strongly correlated in MZ twins than in DZ twins. In that case, lung cancer may appear to be hereditary even when it is not.

In contrast, it may well happen that violation of the equal environment assumption results in an underestimation of the role of genetic factors. The phenomenon of lateral inversion with MZ twins, for example, where some aspect of normal anatomical asymmetry is reversed in one member of a twin-pair, increases variability within these MZ twin-pairs. The common environment is then again, not the same for MZ as for DZ twins and estimation of heritability may not be trusted. Sommer et al. (1999), for example, consider diseases whose pathology is related to cerebral lateralization. The equal environment assumption (1) states that monozygotic twins, like dizygotic twins, share cerebral hemispheric functions. This assumption may be false, since monozygotic twins are more liable to the occurrence of mirror-imaging. They explain the phenomenon as follows. Left-right asymmetry is probably determined as early as the first few cell divisions, long before any morphological sign of asymmetry is visible. The process of twinning in monozygotic twins may interfere with the development of normal lateralization in such a way that for some of the asymmetrical features the resulting twins will not be duplicates, but mirror-images of each other. Mirror-imaging in monozygotic twins has been described for structures that develop from the ectoderm such as hair whorl, eye sight, dentation, neavi and dermatoglyphs, and for handedness. Therefore, it may also be expected for the cerebral hemispheres. Thus, it may also have implications for studies that use twins to investigate the relative contribution of genes and environment in cerebral diseases, such as schizophrenia. In schizophrenia, the left hemisphere is reported to be more affected than the right. A subject with 'mirrored' dominance may become much less disabled by left hemispherical disease processes than a subject with 'standard' dominance. Such unequal involvement of the hemispheres may be relevant in other cerebral diseases such as depression, autism and dyslexia as well.

Despite the above concerns, the twin method remains a useful and common tool in human genetics. It is valid when it is reasonable to believe

that MZ and DZ twins share the same environment with respect to factors that could be linked with the phenotype of interest.

### 2.2.1 Structural equation models

#### General theory

In the literature of twin studies, as well as in psychometrics and econometrics, path-diagrams and structural equation models (SEM) are commonly used to model the effects of exposures on outcomes and to measure the genetic impact on a phenotype (Neale and Cardon, 1992). Roughly, path diagrams are much like causal DAGs (see Chapter 1) in the sense that they represent the data-generating mechanism. This happens by visualizing all causal relations between (measured and unmeasured) variables by means of edges.

Path diagrams incorporate the following conventions (Neale and Cardon, 1992):

- A fundamental distinction is made between independent variables and dependent variables. Independent variables are not caused by other variables in the system.
- Each dependent variable is influenced by an exogenous error term, unless this term is assumed to be (and is fixed to) zero. This error term is a variable which does not correlate with any other determinants of its dependent variable, and which is usually (but not always) uncorrelated with other independent variables.
- Observed variables are enclosed in rectangles. Latent variables are enclosed in ellipses. Error terms are included in the path diagram and are not enclosed.
- A one-way arrow between two variables indicates the possibility of a direct influence of the variable at the tail on the variable at the arrowhead. A two-way arrow between two variables indicates that these variables may be correlated without any assumed direct (causal)

relationship. A two-way arrow from one variable into itself represents the variance of that variable.

- The parameter corresponding to a path is called a path coefficient and the parameter corresponding to a two-way arrow is called a correlation (or covariance) coefficient.
- Coefficients may have two subscripts, the first indicating the variable to which the arrow points, the second showing its origin.

A basic premise of path diagrams is the following:

**Assumption 2** (Causal closure)**.** *All direct influences of one variable to another are included in the path diagram.*

Hence, the non-existence of an arrow between two variables means that these variables are assumed not to be directly related.

Before elaborating on the definition of a path diagram, we give an example of a path diagram in Figure 2.1. The diagram has two dependent variables $\eta_1$ and $\eta_2$, which are affected by 3 independent variables $\xi_1$, $\xi_2$ and $\xi_3$. The three independent variables are correlated with covariance coefficients $(\phi_{12}, \phi_{13}, \phi_{23})$ and variances equal to $\phi_{11}$, $\phi_{22}$ and $\phi_{33}$ respectively. Further, $\zeta_1$ and $\zeta_2$ are the error terms corresponding to $\eta_1$ and $\eta_2$, respectively, and have variances $\psi_{11}$ and $\psi_{22}$ respectively.

Although the description of path diagrams so far is very similar to the description of causal DAGs in Chapter 1, they differ from causal DAGs by the fact that they represent a linear structural equation model. This is a multivariate normal model for the joint distribution of all dependent variables in the path diagram, conditional on the independent variables. Here, this distribution is such that it satisfies all conditional independence relationships implied by the underlying DAG (i.e. the path diagram with bi-directional arrows being replaced by an unmeasured common cause) (Pearl, 2000) and such that all relationships between variables are linear. Because path diagrams represent a linear structural equation model, their formal completeness requires the introduction of error terms unless there is a reason to assume a fully deterministic additive model or unless these are already represented through unmeasured variables in the diagram. Such error

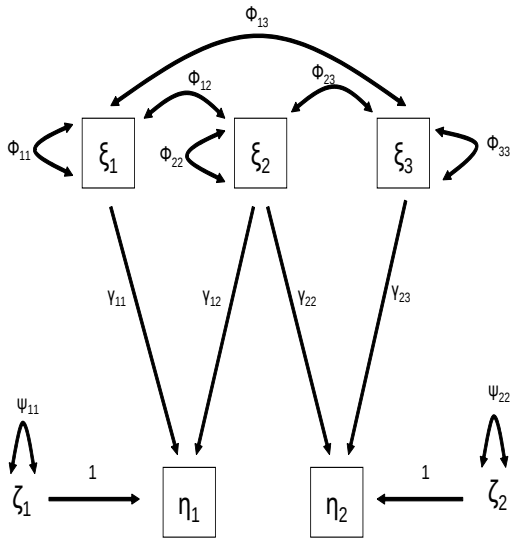


Figure 2.1: *Example of a path diagram*

terms are not usually displayed in causal DAGs, which merely require to include common causes of any 2 variables in the DAG. The other major distinction is that DAGs are inherently non-parametric in the sense of not making any distributional assumptions, whereas path diagrams are fully parametric. This makes path diagrams of more limited usefulness, although the assumption of a linear model may serve as a good approximation for many non-linear functions within limited range.

The structural equation models corresponding to Figure 2.1 are

$$\eta_1 = \alpha_1 + \gamma_{11}\xi_1 + \gamma_{12}\xi_2 + \zeta_1$$

and

$$\eta_2 = \alpha_2 + \gamma_{22}\xi_2 + \gamma_{23}\xi_3 + \zeta_2$$

where $\alpha_1$ and $\alpha_2$ are intercepts and $\zeta_1$ and $\zeta_2$ are independent mean zero, normally distributed variables with variances $\psi_{11}$ and $\psi_{22}$ respectively. We

can rewrite these equations using matrices

$$\begin{pmatrix} \eta_1 \\ \eta_2 \end{pmatrix} = \begin{pmatrix} \alpha_1 \\ \alpha_2 \end{pmatrix} + \begin{pmatrix} \gamma_{11} & \gamma_{12} & 0 \\ 0 & \gamma_{22} & \gamma_{23} \end{pmatrix} \begin{pmatrix} \xi_1 \\ \xi_2 \\ \xi_3 \end{pmatrix} + \begin{pmatrix} \zeta_1 \\ \zeta_2 \end{pmatrix}$$

or, using bold matrix notation,

$$\boldsymbol{\eta} = \boldsymbol{\alpha} + \boldsymbol{\Gamma}\boldsymbol{\xi} + \boldsymbol{\zeta} \tag{2.1}$$

where $\boldsymbol{\eta}$, $\boldsymbol{\alpha}$, $\boldsymbol{\xi}$ and $\boldsymbol{\zeta}$ are column vectors and $\boldsymbol{\Gamma}$ is a matrix of regression coefficients. If we denote $\boldsymbol{\Phi}$ and $\boldsymbol{\Psi}$ for the covariance matrices of $\boldsymbol{\xi}$ and $\boldsymbol{\zeta}$ respectively, we find, using simple matrix algebra, the following expected covariance matrix ($\boldsymbol{\Sigma}$) for $\boldsymbol{\eta}$

$$\boldsymbol{\Sigma} = \boldsymbol{\Gamma}\boldsymbol{\Phi}\boldsymbol{\Gamma}' + \boldsymbol{\Psi} \tag{2.2}$$

$$= \begin{pmatrix} \gamma_{11} & \gamma_{12} & 0 \\ 0 & \gamma_{22} & \gamma_{23} \end{pmatrix} \begin{pmatrix} \phi_{11} & \phi_{12} & \phi_{13} \\ \phi_{12} & \phi_{22} & \phi_{23} \\ \phi_{13} & \phi_{23} & \phi_{33} \end{pmatrix} \begin{pmatrix} \gamma_{11} & 0 \\ \gamma_{12} & \gamma_{22} \\ 0 & \gamma_{23} \end{pmatrix} + \begin{pmatrix} \psi_{11} & 0 \\ 0 & \psi_{22} \end{pmatrix}$$

After collecting data on the observed variables ($\eta_1$, $\eta_2$, $\xi_1$, $\xi_2$ en $\xi_3$), they may be summarized as an observed covariance matrix $\boldsymbol{S}$. This observed covariance matrix is then compared with the expected covariance matrix $\boldsymbol{\Sigma}$ using maximum likelihood, weighted least squares or other estimation methods, to obtain estimates for the unknown parameters in the model. One key issue with structural equation modeling is that it is not always easy to see whether a model or a parameter within a model is identified. Parameters of a model are either overidentified, just identified or underidentified. If all of the parameters fall into the first two classes, the model as a whole is identified, but if one or more parameters are in the third class, the model is not identified. In that case, one must fix the unidentified parameters in the model (e.g. set them equal to zero) until the model becomes identified (Neale and Cardon, 1992), or better, recourse to a sensitivity analysis (Vansteelandt et al. , 2006).

### Application to the twin model

The classical twin model, in which MZ and DZ twins are raised together in the same home, is represented by the path diagram in Figure (2.2). Here, $P_i$, $A_i$, $D_i$, $C_i$ and $E_i$ represent (for both MZ and DZ twins) the observed phenotype ($P_i$) and unobserved additive genetic factors[1] ($A_i$), dominant genetic factors[2] ($D_i$), common environment ($C_i$) and specific (individual) environment ($E_i$) of twin $i$ ($i = 1, 2$).

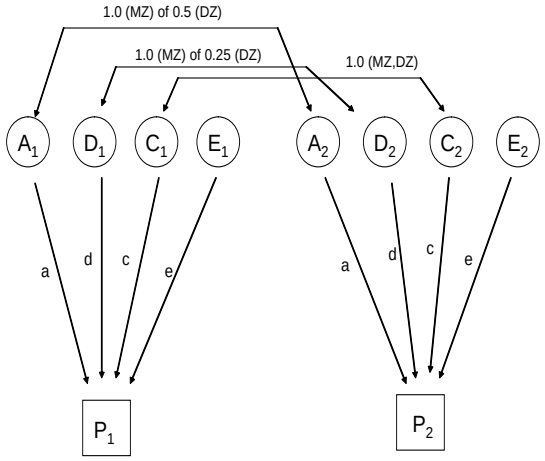


Figure 2.2: *Path diagram of the classical twin model*

Expert knowledge about the correlation between additive and dominant genetic factors, leads to the correlation coefficients shown on the diagram. MZ twins have identical genes, thus, both their additive and dominant genetic factors are perfectly correlated. DZ twins share, on av-

[1]Additive genetic factors refer to a mechanism of quantitative inheritance such that the combined effects of genetic alleles at two or more gene loci are equal to the sum of their individual effects.

[2]Dominance describes a relationship between the effects of different versions of a gene (alleles) on a phenotype. Humans have two copies of each gene, one inherited from each parent. If the combined effect of these two alleles is the same as the effect of having two copies of one of the alleles, we say that allele's effect is dominant over the other.

erage, half their additive and a quarter of their dominant genetic factors (Neale and Cardon, 1992). The common environmental factors of twin 1 and 2 of both MZ and DZ twins, are, by definition, perfectly correlated. The specific environment can be seen as an error term and is assumed uncorrelated with any other variable in the diagram. Since all genetic and environmental factors are unobserved, their variance is assumed to equal 1, without loss of generality. In addition, the path coefficients ($a$, $d$, $c$ and $e$) are assumed to be equal for MZ and DZ twins and for first and second born twins. Intuitively, this is logical for the genetic effects $a$ and $d$ as there is no reason to believe that genes have different effects for MZ twins as for DZ twins. For the environmental effects $c$ and $e$, this follows from Assumption 1 which implies that $c$ and $e$ are the same for MZ and DZ twins. Further, the path diagram 2.2 implicitly assumes

- no genotype-environment correlation, i.e. latent genetic variables $A$ and $D$ are uncorrelated with latent environmental variables $C$ and $E$;

- no genotype $\times$ environment interaction, so that the observed phenotypes are a linear function of the underlying genetic and environmental variables.

Under the assumptions of this path diagram, the goal is now to estimate the impact of genetic factors (i.e. $a$ and $d$) on the phenotype.

Consider a sample of MZ and DZ twin-pairs, ascertained, for example, from a population twin register. Measurements on the phenotype, collected from this sample, are summarized in the covariances, one between measurements of MZ twins and one between measurements of DZ twins, and in terms of the variance of the measurements. To obtain estimates of $a$, $d$, $c$ and $e$, we may compare the empirical covariances with the expected covariance. This expected covariance can be found as in the previous section, using structural equations. The structural equations in this application are

$$\begin{aligned} P_1 &= \alpha + eE_1 + cC_1 + aA_1 + dD_1 \\ P_2 &= \alpha + eE_2 + cC_2 + aA_2 + dD_2 \end{aligned} \tag{2.3}$$

or in matrix notation

$$\begin{pmatrix} P_1 \\ P_2 \end{pmatrix} = \begin{pmatrix} \alpha \\ \alpha \end{pmatrix} + \begin{pmatrix} e & c & a & d & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & e & c & a & d \end{pmatrix} \begin{pmatrix} E_1 \\ C_1 \\ A_1 \\ D_1 \\ E_2 \\ C_2 \\ A_2 \\ D_2 \end{pmatrix}$$

both for MZ and DZ twins. The covariance matrix $\mathbf{\Phi}$ of the independent variables however, is different for MZ and DZ twins:

$$\mathbf{\Phi}_{\mathrm{MZ}} = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 1 \end{pmatrix}$$

and

$$\mathbf{\Phi}_{\mathrm{DZ}} = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0.5 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0.25 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0.5 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0.25 & 0 & 0 & 0 & 1 \end{pmatrix}$$

This leads to the following formulas for the expected covariance of MZ twins ($\mathbf{\Sigma}_{\mathrm{MZ}}$), the expected covariance for DZ twins ($\mathbf{\Sigma}_{\mathrm{DZ}}$) and the variance ($V_{\mathrm{P}}$)

of the phenotype:

$$\boldsymbol{\Sigma}_{\mathrm{MZ}} = \begin{pmatrix} e^2+c^2+a^2+d^2 & c^2+a^2+d^2 \\ c^2+a^2+d^2 & e^2+c^2+a^2+d^2 \end{pmatrix} \tag{2.4}$$

$$\boldsymbol{\Sigma}_{\mathrm{DZ}} = \begin{pmatrix} e^2+c^2+a^2+d^2 & c^2+0.5a^2+0.25d^2 \\ c^2+0.5a^2+0.25d^2 & e^2+c^2+a^2+d^2 \end{pmatrix} \tag{2.5}$$

Comparing these expected (co)variances with the empirical (co)variances, we obtain three estimating equations to estimate 4 unknown parameters:

$$\begin{cases} \mathrm{Cov(MZ)} &= c^2+a^2+d^2 \\ \mathrm{Cov(DZ)} &= c^2+0.5a^2+0.25d^2 \\ V_{\mathrm{P}} &= e^2+c^2+a^2+d^2 \end{cases} \tag{2.6}$$

Therefore, if we are limited to data from a classical twin study, i.e. MZ and DZ twins raised together, it is necessary to impose at least one constraint on the parameters $a$, $c$ or $d$ (for instance, that one of them is zero) to identify the model. Suppose that we have good reasons to believe that $c$ can be ignored. This could be the case, for example, when estimating heritability of late-onset diseases such as Alzheimer's disease. It may then be reasonable to assume that the two members of the twin do not share much environment (any more), that could affect the occurrence of the disease. Then, the equations may be rewritten as

$$\begin{cases} \mathrm{Cov(MZ)} &= a^2+d^2 \\ \mathrm{Cov(DZ)} &= c^2+0.5a^2+0.25d^2 \\ V_{\mathrm{P}} &= e^2+a^2+d^2 \end{cases} \tag{2.7}$$

which leads to identified parameters.

Another, generally superior, approach to resolving the identification problem (Neale and Cardon, 1992) is to collect additional data on, for example, separated MZ twins. Indeed, the covariance (Cov(MZA)) between the phenotypes obtained from these twins leads a fourth estimating equation

$$\mathrm{Cov(MZA)} = a^2+d^2$$

which can be added to (2.6) to make all 4 parameters identified. Additional information on 'normal' siblings does not resolve the identification problem, since these share the same proportion of genes like DZ (fraternal) twins. Measurements on half siblings or cousins, on the other hand, do add information since the expected covariances between these are $0.25a^2$ and $0.125a^2$, respectively (Neale and Cardon, 1992).

### 2.2.2 Random effects models

We will now describe an alternative method for testing and estimating heritability, introduced by Sham (1998), based on random intercept models with exchangeable correlation structure. The method is mathematically equivalent to the one described in the previous section.

Consider a sample of MZ (or DZ) twin-pairs in which the phenotype $P$ has been measured on each individual. Let the value of the phenotype for twin $j$ ($j = 1, 2$) in pair $i$ ($i = 1, ..., n$) be $p_{ij}$. In random intercept models with exchangeable correlation structure it is assumed that the outcome variable $P$ is determined by two underlying variables, say $B$ and $W$, in which $B$ is perfectly correlated between members of the same twin-pair but uncorrelated between members of different twin-pairs, while $W$ is uncorrelated between any two individuals. That is, assume that

$$P_{ij} = B_i + W_{ij} \tag{2.8}$$

with $W_{ij}$, $i = 1, ..., n, j = 1, 2$ and $B_i$, $i = 1, ..., n$ mutually independent mean zero normally distributed variates. Assume that this model holds for both MZ and DZ twins, with possibly different variance components (i.e. variance and covariance). Here, the variable $B$ contributes to the variation between, but not within twin-pairs, whereas the variable $W$ contributes to the variation between individuals, both between and within twin-pairs.

#### Testing for heritability

A test for heritability is based on the comparison of intraclass correlations between MZ and DZ twins. An intraclass correlation is the correlation between two individuals in the same class (i.e. the same twin-pair).

It can be calculated based on a one-way ANOVA or based on the variance-components estimates of the random intercept model introduced above. Two assumptions need to hold in order to yield the intraclass correlation.

**Assumption 3.** *The variance $V_P$ of the trait is equal for MZ and DZ twins.*

and

**Assumption 4.** *The variation within a twin-pair is not related to the average of the pair (i.e. homoscedasticity).*

Assumption 3 may be violated for several reasons. It is possible, for example, that DZ twins with extreme values of the traits are less likely to be sampled and that this is not the case for MZ twins. This, however, does not happen when using data from a twin register. Alternatively, there may be factors that operate on MZ but not DZ twins (or vice versa), e.g. in DZ twins there is more differential fertility of the mother than in MZ twins; such factors will contribute variability to one type of twins but not the other. Yet another possibility is reciprocal twin-interaction, where the trait value of one twin has a direct effect on the trait value of the other twin. This interaction can be cooperative (high value in one twin increases the value of the other twin) or competitive (high value in one twin decreases the value of the other twin). For twins who deviate in the same direction from the overall population mean, cooperative interaction tends to increase their distance from this mean. For twins deviating in opposite directions from the overall population mean, cooperative interaction tends to decrease their distance from this mean. Since deviation of both twins in the same direction is more frequent for MZ than for DZ twins (assuming that genetic factors are operating), cooperative twin-interaction leads to a greater trait variance in MZ than in DZ twins. Competitive interaction has the opposite effect, i.e. a smaller trait variance in MZ than in DZ twins.

With concern for violation of assumption 3, one may test it via the $F$-statistic

$$F = \frac{\mathrm{MST}_{\mathrm{MZ}}}{\mathrm{MST}_{\mathrm{DZ}}}$$

where $\text{MST}_{\text{MZ}}$ and $\text{MST}_{\text{DZ}}$ are the total mean squares for MZ and DZ twins respectively. These are obtained by dividing the total sum of squares ($\text{SST}_{\text{MZ}}$ and $\text{SST}_{\text{DZ}}$) by their respective degrees of freedom $2n_{\text{MZ}} - 1$ and $2n_{\text{DZ}} - 1$, with $n_{\text{MZ}}$ the number of MZ twin pairs and $2n_{\text{DZ}} - 1$ the number of DZ twin pairs. The total sum of squares can be calculated for MZ and DZ twins separately and equals $\sum_i \sum_j (p_{ij} - p_{..})^2$, where $p_{..} = (\sum_i \sum_j p_{ij})/(2n)$ with $n = n_{\text{MZ}}$ and $n = n_{\text{DZ}}$ for MZ and DZ twins respectively. The test statistic $F$ follows an $F$-distribution with $(2n_{\text{MZ}} - 1,\ 2n_{\text{DZ}} - 1)$ degrees of freedom under the hypothesis of equal variances.

Assumption 4 can be tested by means of two tests for homoscedasticity. Heteroscedasticity arises when the effect sizes of unshared factors (i.e. factors that cause variation within pairs) in a twin-pair depend on the level of the shared factors present in the pair. The variation within a twin-pair can be measured by the absolute pair-difference, defined as $p_{i-} = |p_{i1} - p_{i2}|$. Denoting the average absolute pair-difference of the sample as $p_{.-}$, the product moment correlation between pair-means and absolute pair-difference is

$$d = \frac{\sum_i (p_{i.} - p_{..})(p_{i-} - p_{.-})}{(\sum_i (p_{i.} - p_{..})^2 \sum_i (p_{i-} - p_{.-})^2)^{1/2}}$$

The significance of this correlation can be tested using Fisher's $z$ transformation

$$z = \frac{1}{2} \ln \left[ \frac{1 + d}{1 - d} \right] \tag{2.9}$$

which is normally distributed in large samples under the null hypothesis of no correlation, with mean zero and variance equal to $1/(n - 3)$ (again, $n = n_{\text{MZ}}$ and $n = n_{\text{DZ}}$ for MZ and DZ twins respectively).

When these assumptions are justified, the intraclass correlation can be calculated as follows. It follows from the random intercept model, that the covariance between the outcomes of two individuals in the same twin-pair is $\text{Var}(B)$, and the variance of each individual's outcome is $\text{Var}(B) + \text{Var}(W)$. Thus, the intraclass correlation is

$$\rho = \frac{\text{Var}(B)}{\text{Var}(B) + \text{Var}(W)} \tag{2.10}$$

Estimates for these variance components (for MZ and DZ twins separately) can be obtained from the estimates of the variance of the random intercept (i.e. $B$) and the variance of the residual error (i.e. W) in the random intercept model for $P$. We then obtain estimates $r_{\mathrm{MZ}}$ and $r_{\mathrm{DZ}}$ for the intraclass correlation of MZ and DZ twins, respectively. The hypothesis of a genetic contribution to $P$ predicts a greater value for $r_{\mathrm{MZ}}$ than for $r_{\mathrm{DZ}}$. Using Fisher's $z$-transformation

$$z = \frac{1}{2}\ln\left[\frac{1+r}{1-r}\right] \tag{2.11}$$

yields $z_{\mathrm{MZ}}$ and $z_{\mathrm{DZ}}$ when using $r = r_{\mathrm{MZ}}$ and $r = r_{\mathrm{DZ}}$, respectively. A formal test statistic is then

$$z_{\mathrm{T}} = \frac{z_{\mathrm{MZ}} - z_{\mathrm{DZ}}}{\left(\frac{1}{n_{\mathrm{MZ}}-2} + \frac{1}{n_{\mathrm{DZ}}-2}\right)^{1/2}} \tag{2.12}$$

which follows approximately a standard normal distribution in large samples under the null hypothesis of equal intraclass correlation for MZ and DZ twins.

### Estimating heritability

If there is evidence of a greater intraclass correlation among MZ than among DZ twins, then one may wish to proceed to obtain some measure of the relative importance of genetic to environmental factors, i.e. heritability.

There are two different such measures that are both often referred to as heritability. The first is called broad heritability and is the proportion of total phenotypic variance accounted for by all genetic components (i.e. additive and dominant). The second is called narrow heritability, or just heritability, and is the proportion of phenotypic variance accounted for by the additive genetic component.

As already mentioned in the previous section, the classical twin method is based on the partition of genetic variance into additive and dominance components, and the partition of environmental variance into shared and non-shared components. It assumes a partly common environment for the

twin-pairs, meaning that one can define a single component of variance, $V_C$, to represent the common environmental variance and a single component of variance $V_E$, to represent the remaining, non-shared, (individual) environmental variance. Assumption 1 implies that this common and individual environment variability is equal for MZ and DZ twins. The estimation of heritability is based on modeling the total phenotypic variance, $V_P$, both for MZ and DZ twins, as

$$V_P = V_A + V_D + V_C + V_E$$

in which $V_A$ and $V_D$ represent to variability due to additive and dominant genetic components respectively. Note that this decomposition ignores epistatic genetic effects[3] and assumes no gene $\times$ environment interaction. Thus, the model is based on the same assumptions implied by the path diagram of the SEM method in the previous section.

Under these assumptions, an estimate for heritability can be obtained as follows. The relationships between the variance components ($V_A$, $V_D$, $V_C$, $V_E$) and the intraclass correlations for MZ and DZ twins are (Sham, 1998)

$$\begin{aligned} \rho_{\text{MZ}} &= \frac{V_A + V_D + V_C}{V_A + V_D + V_C + V_E} \\ \rho_{\text{DZ}} &= \frac{1/2V_A + 1/4V_D + V_C}{V_A + V_D + V_C + V_E} \end{aligned} \tag{2.13}$$

Estimates $(r_{\text{MZ}}, r_{\text{DZ}})$ for these intraclass correlations can be obtained using the method described in the previous section, based on the random intercept model (2.8). Since we are interested in the proportions of the variance components, we can set the total phenotypic variance, $V_P$, to 1 with no loss of generality. However, even when doing so, estimating the values of three unknown parameters ($V_A$, $V_D$, $V_C$) by equating them with the two sample correlations $(r_{\text{MZ}}, r_{\text{DZ}})$ yields no unique solution.

To gain insight, note that the above model requires that the ratio of $\rho_{\text{MZ}}$ and $\rho_{\text{DZ}}$ (the true correlations) takes values between 1 (when $V_C > 0$,

[3] that is, interactions between additive and dominant genetic factors.

$V_A = V_D = 0$) and 4 (when $V_D > 0$, $V_A = V_C = 0$). Moreover, $\rho_{\text{MZ}}/\rho_{\text{DZ}} = 2$ when $V_A > 0$ and $V_C = V_D = 0$, $\rho_{\text{MZ}}/\rho_{\text{DZ}} < 2$ when $V_A > 0$, $V_C > 0$ and $V_D = 0$, and $\rho_{\text{MZ}}/\rho_{\text{DZ}} > 2$ when $V_A > 0$, $V_D > 0$ and $V_C = 0$. Hence, when $r_{\text{MZ}}/r_{\text{DZ}} < 1$ or $r_{\text{MZ}}/r_{\text{DZ}} > 4$, we conclude that the model is inappropriate. When $1 \leq r_{\text{MZ}}/r_{\text{DZ}} \leq 2$, we may be willing to assume that $V_D = 0$ and then estimate $V_A$ and $V_C$ as

$$\begin{aligned} \hat{V}_A &= 2r_{\text{MZ}} - 2r_{\text{DZ}} \\ \hat{V}_C &= 2r_{\text{DZ}} - r_{\text{MZ}} \end{aligned} \tag{2.14}$$

In this case, we cannot obtain an estimate of broad heritability. The estimate of narrow heritability is then

$$H_N^2 = \hat{V}_A$$

When $2 < r_{\text{MZ}}/r_{\text{DZ}} \leq 4$, we set $V_C = 0$ and estimate $V_A$ and $V_D$ as

$$\begin{aligned} \hat{V}_A &= 4r_{\text{DZ}} - r_{\text{MZ}} \\ \hat{V}_D &= 2r_{\text{MZ}} - 4r_{\text{DZ}} \end{aligned} \tag{2.15}$$

An estimate of broad heritability is then

$$H_B^2 = \hat{V}_A + \hat{V}_D$$

and an estimate of narrow heritability is the same as in the previous case.

This procedure does not imply that $V_C$ and $V_D$ cannot coexist, but merely that they cannot be jointly estimated with the data available.

Since these heritability estimates are linear combinations of intraclass correlations estimates, their approximate standard errors can be obtained by considering the sampling variances of the intraclass correlations (Sham, 1998). The variance of the intraclass correlation, $r$, estimated from data on $n$ twin-pairs is approximately

$$Var(r) = \frac{(1-\rho^2)^2}{n}$$

where $\rho$ is the true intraclass correlation. When $1 < r_{\text{MZ}}/r_{\text{DZ}} \leq 2$, for example, narrow heritability is estimated by $H_N^2 = \hat{V}_A = 2r_{\text{MZ}} - 2r_{\text{DZ}}$ so that its standard error is approximately

$$SE(h^2) = 2\left(\frac{(1-r_{\text{MZ}}^2)^2}{n_{\text{MZ}}} + \frac{(1-r_{\text{DZ}}^2)^2}{n_{\text{DZ}}}\right)^{1/2}$$

This method can be used to obtain approximate standard errors of $h^2$ and $H^2$ for the different ranges of values of the $r_{\text{MZ}}/r_{\text{DZ}}$ ratio.

### 2.2.3 Heritability of birth weight in the East Flanders Prospective Twin Survey

The East Flanders Prospective Twin Survey (EFPTS) (Loos et al., 1998; Derom and Derom, 2005) gathers data on all twins born in East Flanders since 1976. This register is renowned internationally because it concerns a population and because it is the only twin register which gathers detailed information on zygosity. Zygosity is determined through sequential analysis of fetal sex, fetal membranes, and umbilical cord blood groups and by DNA fingerprinting based on allelic similarity within a twin pair of short tandem repeat loci on nine different chromosomes. Overall, zygosity (and chorionicity) are determined with an accuracy of over 99%.

Using data from the EFPTS on twins born between 1976 and 2002, we will examine the heritability of birth weight. Before we calculate the intraclass correlations for MZ and DZ twins, we need to ascertain whether Assumptions 3 and 4 are met. There are 1610 MZ twin pairs and 3290 DZ twin pairs. The total mean squares (MST) for MZ and DZ twins can be found in Table 2.1 in the third row. Using an F-test (see previous section), we test whether Assumption 3 is met, i.e. whether birth weights are equally variable in MZ and DZ twins and find a $p$-value of 0.9999992, suggesting no evidence to reject the hypothesis of equal variances.

We test the homoscedasticity (Assumption 4) for MZ and DZ twins by calculating the absolute pair-differences, using Fisher's $z$ transformation and comparing the obtained statistic to a normal distribution with mean zero and variance equal to $1/(n-3)$ ($n = n_{\text{MZ}}$ and $n = n_{\text{DZ}}$ for MZ and

| | MZ twins | DZ twins |
|---|---|---|
| $n$ | 1610 | 3290 |
| SST | 1 130 891 100 | 1 999 473 503 |
| MST | 351 317.5 | 303 917.5 |
| Var($B$) | 275 644 | 210 575 |
| Var($W$) | 76 313 | 93 627 |
| $r$ | 0.7832 | 0.6922 |
| $p$-value of $z$-test for heritability | $< 0.0001$ | |

Table 2.1: *Calculations to obtain estimated intraclass correlation (r) for MZ and DZ twins*

DZ twins respectively). The results can be found in Table 2.2. The obtained $p$-value for MZ twins allow us to conclude that there is no evidence of heteroscedasticity. For DZ twins however, there is some evidence of dependency between the pair-differences and pair means. However, we note that the DZ twin sample is very large, which increases the chance of finding significance. Moreover, the upper bound of the 95% confidence interval is close to zero. Thus, this violation of the assumption will not have a major impact on the heritability estimate.

| | MZ twins | DZ twins |
|---|---|---|
| $p_{.-}$ | 282.03 | 324.53 |
| $d$ | 0.0263 | 0.0451 |
| $z$ | 0.0263 | 0.0451 |
| $p$-value | 0.28 | 0.0088 |
| 95% CI for $z$ | [-0.023;0.075] | [0.011;0.079] |

Table 2.2: *Calculations to test the Assumption 4 for MZ and DZ twins. CI= confidence interval*

Now, we fit a random intercept model for birth weight using proc mixed in SAS (version 9.1) with exchangeable correlation structure and different covariance parameters for MZ and DZ twins. We obtain the estimates for these covariance parameters in the fourth and fifth row of Table 2.1. Us-

ing these estimates, the intraclass correlation of MZ and DZ twins can be calculated (last row of Table 2.1). We find that there is a difference in intraclass correlation between MZ and DZ twins, which is an indication that birth weight might be subject to genetic influences. This is confirmed by testing the hypothesis of equal intraclass correlations based on the $z_T$ test statistic (2.12) which gives a $p$-value smaller than 0.0001. To estimate heritability, we first calculate the ratio of the two intraclass correlations, i.e. 0.7832/0.6922=1.1315 which lies between 1 and 2. Thus, we assume that the variability due to dominant genetic factors can be ignored and estimate the additive genetic variability ($V_A$) and the common environmental variability ($V_C$) as

$$\begin{aligned} \hat{V}_A &= 2r_{\text{MZ}} - 2r_{\text{DZ}} = 0.182 \\ \hat{V}_C &= 2r_{\text{DZ}} - r_{\text{MZ}} = 0.6012 \end{aligned} \tag{2.16}$$

This leads to an estimate of narrow heritability of 0.182 with a standard error of 0.026 and a 95% confidence interval of $[0.13; 0.23]$. This means that 18.2% of the total phenotypic variability is explained by additive genetic factors. Note that, if we would have assumed that variability due to common environmental factors could be ignored, we would have obtained an estimate of 1.99 for narrow heritability, which is not interpretable.

### 2.2.4 Heritability and confounding

Estimation of heritability brings along the question whether heritability estimates may be confounded by extraneous factors. In particular, it is of interest to assess whether standard twin analyses could systematically find a given phenotype to be heritable, even when in fact it is not at all related to genetic causes. To the best of our knowledge, this problem has only been tangentially addressed in the literature (Neale and Cardon, 1992). Common practice is to adjust the analyses for factors such as age. From the perspective of confounding adjustment, this approach will not remove confounding bias. Instead it will remove part of the variability due to environmental factors and thus will yield larger heritability estimates by the fact that they pertain to more homogeneous subpopulations.

To gain insight into the problem of confounding, consider the causal DAG underlying the standard path diagram for twin analysis (see Figure 2.2). It is then natural to think that heritability estimates will be confounded whenever there exist common causes of the phenotype, the genetic and the environmental factors (e.g. age of first pregnancy, as it is related to adverse perinatal outcomes and may also be genetically determined). However, this interpretation is misleading. This is because estimation of heritability is based on comparing intraclass correlations of the phenotype between MZ and DZ twins (and not by using actual data on genetic or environmental factors). Confounding of heritability estimates will therefore arise whenever the phenotypic correlation differs between MZ and DZ twins, even in the absence of a genetic effect. We thus conclude that heritability estimates may be confounded whenever the association between zygosity and phenotype is confounded. This is represented in the diagram of Figure 2.3. Note that, since DAGs are inherently non-parametric, standard d-separation rules continue to apply, even though this association is now represented as a difference in phenotypic correlation. It follows that heritability estimates are confounded whenever there exist common causes of zygosity and the phenotype.

**Data analysis**

When estimating the heritability of birth weight, one such common cause of zygosity and birth weight is the type of conception. Indeed, it has been established that the type of conception (spontaneous or not) affects birth weight (Verstraelen et al., 2005). Furthermore, until recently, when artificial reproductive technologies like in vitro fertilization (IVF) or intracytoplasmic sperm injection (ICSI) were used to help conception, often more than one embryo was placed back into the womb. This hence implied a bigger chance of having a DZ twin than in spontaneously conceived twins. Another possible confounder for the effect of zygosity on birth weight is maternal age. The older the mother at the time of conception, the more chance of having a DZ twin and the more risk of a low birth weight baby. Applying d-separation teaches us that the causal effect of zygosity on birth weight can be obtained by adjusting the analysis for type of conception and

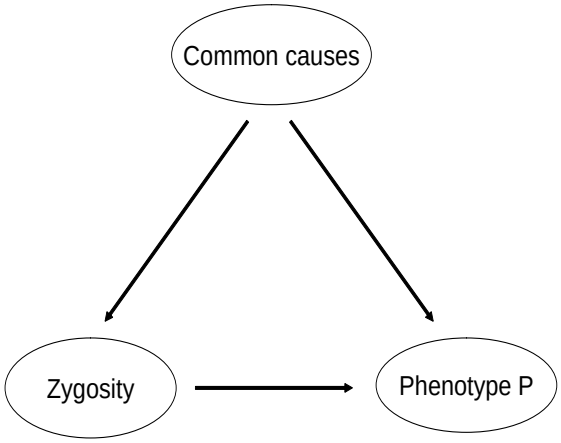


Figure 2.3: *Causal diagram for estimating heritability of a phenotype $P$*

maternal age. Because type of conception and maternal age are amongst the few factors that might affect the prevalence of MZ versus DZ twins, confounding poses no major issues when estimating the heritability of a given trait, provided that these factors are adjusted for. For the EFPTS register, adjustment for type of conception and maternal age does not modify the heritability estimate, thus the estimated heritability in Section 2.2.3 is not subject to confounding by type of conception and maternal age.

### Simulation study

We illustrate that the heritability estimate for a phenotype may be confounded when zygosity and the phenotype share a common cause, via a simple simulation study. We generate a twin data set according to the DAG in Figure 2.3, but with no heritability. We generate a binary common cause $C$ (e.g. type of conception; spontaneous or not) with 30% chance of 'success' and a binary variable representing zygosity with a $20 + 75C$ % chance of a dizygotic twin. The phenotype $P$ (e.g. IQ) for the twins is generated using a random intercept model with residual standard deviation

equal to 3 and with mean $180+12C+b_i$ in which $b_i$ $(i = 1, 2)$ represents the random intercept which is mean zero normally distributed with standard deviation equal to 4. Note that the phenotype is generated independently of zygosity to reflect the fact that there is no heritability.

The results can be viewed in Table 2.3. We find that the estimated heritability without adjusting for type of conception is significantly different from zero. There is a substantial bias of 0.30. After adjusting the analysis for type of conception, we find that heritability is no longer significant. We conclude that estimating heritability without adjusting for the common causes may give seriously misleading results.

| | No adjustment | Adjustment for $C$ | |
|---|---|---|---|
| | | $C = 0$ | $C = 1$ |
| ITC for Z=MZ | 0.84 | 0.644 | 0.64 |
| ITC for Z=DZ | 0.68 | 0.6399 | 0.6396 |
| heritability | 0.31 | 0.0082 | 0.0009 |
| SE | 0.0053 | 0.0075 | 0.0076 |
| 95% CI | [0.30;0.32] | [-0.0065;0.023] | [-0.014;0.016] |

Table 2.3: *Intratwin correlations (ITC) for monozygotic (MZ) and dizygotic (DZ) twins and estimation of heritability with standard error (SE) and 95% confidence interval (CI)*

## 2.3 Estimation of causal exposure effects based on twin data

Twin data are not only useful for estimating genetic effects, they also have a rich structure for inferring causal effects because the comparability of twin children can be exploited to obtain effect estimates that are consistent in the presence of unmeasured confounders that are constant within twins, e.g. parental characteristics, environmental factors,...

For example, when estimating the effect of smoking on lung cancer, twins offer an estimator that is 'protected' against factors like smoking by

friends (which are often the same for both twin children), smoking by parents,....These factors are often unmeasured in studies about the effects of smoking, leading to biased estimates. A study using twin data and making comparisons within twins, would not encounter this problem (Carmelli and Page, 1996). Carlin et al. (2005) examined the effect of cord blood erythropoietin ($X$) on birth weight ($y$) in twin data. They describe 3 methods to analyse the data, two of them being based on the principle of making comparisons within twins. These methods thus protect the estimated effect against confounders that are common for both twin children (e.g. maternal factors such as diet and socioeconomic background). The first method is developed by Neuhaus and Kalbfleisch (1998) and models the expected outcome as follows

$$E(Y_{ij}) = \beta_0 + \beta_w(X_{ij} - \overline{X}_i) + \beta_b\overline{X}_i$$

where $\overline{X}_i$ represents the mean value of $X$ for twin pair $i$. The exposure effect is separated into a within- ($\beta_w$) and between- ($\beta_b$) twin component. The model is fitted using a method that respects the paired structure of the data, such as mixed models (Verbeke and Molenberghs, 1997, 2000) or generalized estimating equations (Diggle et al., 1994). The model is commonly used for general clustered data, however, its validity and efficiency has not been formally studied.

The second proposed method is based on analyzing paired-difference values, where differences between $X$ and $Y$ values within each pair are defined by ordening the twins according to birth order, leading to $D_i^Y = Y_{i1} - Y_{i2}$ and $D_i^X = X_{i1} - X_{i2}$. The method then models these transformed values as

$$E(D_i^Y) = \beta D_i^X$$

and uses mixed models or generalized estimating equations to obtain a 'difference-in-difference' estimator for $\beta$. Carlin et al. (2005) show that $\beta$ represents the same within-cluster effect as $\beta_w$ in the model of Neuhaus and Kalbfleisch (1994).

In Chapter 3, we develop a general methodology for clustered data with arbitrary correlation structure based on the principle of making comparisons within clusters. In particular, we develop semi-parametric efficient

estimators for the parameters indexing marginal linear and loglinear models which include unmeasured confounders that are constant within clusters. On the basis of the resulting 'conditional generalized estimating equations', we study the validity of the adjustment procedure proposed by Neuhaus and Kalbfleisch (1998). By construction, other common types of estimators that offer protection against unmeasured cluster-level confounders can be viewed as estimators within our class, e.g. conditional likelihood estimators (Diggle, Liang and Zeger, 1994; Verbeke, Spiessens and Lesaffre, 2001), difference-in-difference estimators (Abadie, 2005; Carlin et al., 2005),... Our approach extends these to general covariance structures and nonlinear link functions under weaker assumptions.

# Chapter 3

# Conditional generalized estimating equations for the analysis of clustered and longitudinal data


### Summary

A common and important problem in clustered sampling designs is that the effect of within-cluster exposures (i.e. exposures that vary within clusters) on outcome may be confounded by both measured and unmeasured cluster-level factors (i.e. measurements that do not vary within clusters). When some of these are ill/not accounted for, estimation of this effect through population-averaged models or random-effects models may introduce bias. We accommodate this by developing a general theory for the analysis of clustered data which enables consistent and asymptotically normal (CAN) estimation of the effects of within-cluster exposures in the presence of cluster-level confounders. Semi-parametric efficient estimators are obtained by solving so-called conditional generalized estimating equations (CGEE). In this chapter, we compare this approach with a popular proposal by Neuhaus and Kalbfleisch (1998) who separate the exposure effect into a within- and between-cluster component within a random intercept model. We find that the latter approach yields consistent and efficient estimators when the model is linear, but is less flexible in terms of model specification. Under nonlinear models, this approach may yield inconsistent

and inefficient estimators, though with little bias in most practical settings.

[Original co-author: S. Vansteelandt]

## 3.1 Introduction

Clustered data, such as arise in twin studies, multicenter studies, longitudinal studies, etc., are frequently encountered in practice. They are notable for allowing improved inference for the causal effect of within-cluster exposures on outcome by providing naturally matched subjects. In particular, comparisons of subjects within clusters yield exposure effects that are protected against cluster-level confounders.

Methods for clustered data differ in the way how they infer exposure effects: through within-cluster comparisons only, through between-cluster comparisons only, or through a combination of both. Standard inference for population-averaged models or random-effects models (which implicitly assume independence of cluster effects and covariates) usually relies on information obtained from both within- and between-cluster comparisons. Estimation of the effect of within-cluster exposures on outcome can therefore be seriously misleading under these models whenever important cluster-level confounders are unmeasured or ill accounted for (Neuhaus and Kalbfleisch, 1998; Ten Have et al., 2004; Palta and Yao, 1991; Chao et al., 1997). In view of this, considerable attention has been devoted to statistical methods which infer exposure effects solely through within-cluster comparisons and hence yield valid estimates even in the presence of cluster-level confounders (see for example Begg and Parides, 2003; Berlin et al., 1999; Diggle, Liang, and Zeger, 1994; Neuhaus and Kalbfleisch, 1998; Ten Have et al., 2004; Verbeke, Spiessens, and Lesaffre, 2001), though at the expense of some efficiency loss (Palta and Yao, 1991; Chao et al., 1997). Among these, conditional likelihood methods (Diggle et al., 1994, Verbeke et al., 2001) are popular when outcomes are dichotomous and independent within clusters. For continuous outcomes which lend themselves to linear modeling, such methods have been used to allow inference for within-cluster expo-

sures in linear mixed models with uncorrelated residual errors (Verbeke et al., 2001). However, a possible drawback of conditional likelihood methods which has been repeatedly suggested in the literature, is that in addition to removing the effects of cluster-level confounders, they also remove the effects of cluster-level covariates that may be of interest. Neuhaus and Kalbfleisch (1998) overcome this by separating exposure effects into effects that are shared among cluster members and subject-specific effects. Specifically, they consider generalized linear mixed-effects models with conditional mean

$$E(Y_{ij}|b_i, \mathbf{X}_i) = h\left\{\beta_0 + b_i + \beta_w(X_{ij} - \bar{X}_i) + \beta_b \bar{X}_i\right\} \tag{3.1}$$

and independent residuals, conditional on $b_i$ and $\mathbf{X}_i$, where $Y_{ij}$ and $X_{ij}$ are the outcome and exposure for the $j$th subject ($j = 1, ..., n_i$) in the $i$th cluster ($i = 1, ..., n$), $\mathbf{X}_i = (X_{i1}, ..., X_{in_i})'$, $\bar{X}_i$ is the average exposure in the $i$th cluster, $h$ is a known inverse link function and $b_i$ is a random intercept (assumed to be independent of $\mathbf{X}_i$). They show by example that for $h$ the identity or inverse logit link, ordinary mixed-effects estimation of the exposure effect $\beta_w$ yields an estimate that is nearly identical to the one obtained via conditional likelihood estimation. The simplicity of this approach, along with its potential to yield estimates of between-cluster effects, has made it a popular alternative to conditional likelihood methods (Begg and Parides, 2003; Ten Have et al., 2004).

In this article, we study the validity of the proposal by Neuhaus and Kalbfleisch (1998) for inferring within-cluster exposure effects under standard conditional mean models which may involve unmeasured cluster-level confounders, nonlinear link functions and general covariance structures. Our focus on such general models is motivated by the fact that, as we argue in Section 3.4, models which involve within- and between-cluster exposure effects cannot be viewed as data-generating models and have limited flexibility.

We start in Section 3.3 by developing a class of estimators which contains (up to asymptotic equivalence) all consistent and asymptotically normal (CAN) estimators for the within-cluster exposure effects indexing general conditional mean models. This class thus unites the different existing

analysis methods, such as regression of change in response on change in covariates (Louis et al., 1986). In addition, we identify a (locally) efficient estimator within our class and thus a recommended analysis method. The proposed estimators are obtained by solving unbiased estimating equations, which we call conditional generalized estimating equations (CGEE) to express that they form a semi-parametric alternative to conditional likelihood estimation with the flexibility and properties of generalized estimating equations (Liang and Zeger, 1986). Informally, CGEE remove any dependence on the cluster-level part of the model by making within-cluster comparisons. The corresponding estimators thus remain CAN even when there are unmeasured cluster-level confounders and/or when the cluster-level part of the model is misspecified.

In Section 3.4, we compare CGEE estimators with those obtained by separating exposure effects into within- and between-cluster effects (Neuhaus and Kalbfleisch, 1998). We find that, under our model with the identity link, the latter estimators are asymptotically equivalent to CGEE estimators within our class, thus confirming their validity. However, we show that using this approach may lead to inconsistent and inefficient estimators of the within-cluster effects under models with log link, in which case consistent and efficient estimators can be obtained via CGEE. This is confirmed via simulation studies in Section 3.6, where we find relatively weak bias in most practical settings, but a more important loss of efficiency. CGEE has the drawback of using computationally slightly more complex estimators. It has the advantage of yielding efficient and consistent estimators in both linear and loglinear marginal models with general covariance structures and not being restricted to limited and more difficult-to-interpret models (namely, those that involve within- and between-cluster effects).

## 3.2 Clinical effect of imipramine on depression

We consider data from a longitudinal psychiatric study to examine the clinical effects of imipramine (IMI) on depression (Reisby et al., 1977). Sixty six depressed inpatients were enrolled in the study and baseline characteristics were recorded (including gender and diagnosis of endogenous

depression). After a one week run-in placebo period, they were given the same daily dose of 225 mg IMI during four weeks. Desipramine (DMI) is the active 'in vivo' metabolite of IMI and, as such, is responsible for the antidepressant effects of IMI. The plasma levels of both IMI and DMI were therefore measured at the end of every treatment week and the log concentration ratio of DMI and IMI was considered as a measure for IMI absorption, large values being indicative of an effective absorption of IMI into the blood. In addition, the Hamilton depression rating scale (HDRS) score was measured at the start and end of the run-in period, as well as at the end of every treatment week. The change in HDRS score (i.e. since baseline) was reported as the outcome. Positive outcomes are indicative of an improved depression status.

Our goal is to estimate the effect of IMI absorption on the expected change in HDRS score. Standard inference under population-averaged models or random-effects models might lead to biased estimates of this effect whenever, as often, (a) there are unmeasured prognostic factors for depression which are associated with IMI absorption (e.g. body size, use of other medicines, amount of sleep); or (b) the effect of these measured prognostic factors was not adequately modeled. In this chapter, we develop a flexible strategy which protects against unmeasured or misspecified cluster-level confounders by making efficient within-patient comparisons under limited modeling assumptions.

## 3.3 Conditional generalized estimating equations

The study design that we consider, collects data $(\mathbf{Y}_i, \mathbf{X}_i, \mathbf{V}_i, S_i)$ for each of $i = 1, ..., n$ independent clusters/patients. Here, $\mathbf{Y}_i = (Y_{i1}, ..., Y_{in_i})'$ is a vector of outcome measurements (e.g. depression change) for each of $j = 1, ..., n_i$ subjects/occasions, $\mathbf{X}_i = (X_{i1}, ..., X_{in_i})'$ is a vector of primary exposure measurements (e.g. log concentration ratio of DMI and IMI), $\mathbf{V}_i = (V_{i1}, ..., V_{in_i})'$ is a remaining within-cluster variable (e.g. time, or a within-cluster confounder) and $S_i$ contains cluster-level measurements (e.g. gender, diagnosis of endogenous depression at baseline, ...). With a slight abuse of notation, we allow $X_{ij}$, $V_{ij}$ and $S_i$ to be vector measurements. The goal of the study is to infer the effect $\beta$ of exposure $X_{ij}$ on expected outcome in the conditional mean model

$$E(Y_{ij}|\mathbf{X}_i, \mathbf{V}_i, S_i, b_i) = h(\alpha + X_{ij}\beta + V_{ij}\gamma + S_i\delta + b_i), \tag{3.2}$$

where $h(.)$ is the identity or inverse log link and $b_i$ is an unmeasured cluster-level variable. Unlike in ordinary random effects models (Diggle et al., 1994; Ten Have et al., 2004), $b_i$ is allowed to be correlated with the remaining variables in the model and thus allows for unmeasured cluster-level confounders and/or misspecified effects of $S_i$ on outcome. Note that formulation (3.2) is flexible in that it makes no distributional assumptions (in particular, it allows arbitrary covariance structure) and that it allows interactions between cluster-level variables (by including these in $S_i$), between cluster-level variables and $b_i$ (by including these in $b_i$), between within-cluster variables and between within-cluster and cluster-level variables (by including these in $V_{ij}$). In the context of longitudinal studies, it allows, for example, for time effects (by including these in $V_{ij}$) and adjustment for previous exposures (by letting these be part of $X_{ij}$).

Our goal is to conduct inference for the association $\boldsymbol{\omega} = (\beta, \gamma)'$ of within-cluster variables $\mathbf{L}_{ij} = (X_{ij}, V_{ij})$ with outcome under the observed data model defined by (3.2). This is challenging because the linear predictor includes an infinite-dimensional nuisance parameter $b_i$. Using semiparametric theory (Bickel et al., 1993), we derive, up to asymptotic equivalence, the set of all CAN estimators for $\boldsymbol{\omega}$ under this model. Theorem 3

states the main results and motivates our construction of estimators for $\boldsymbol{\omega}$ in this model.

Suppose first that $h(.)$ is the identity link. Part 1a of Theorem 3 then shows that CAN estimators for $\boldsymbol{\omega}$ under model (3.2) can be obtained by solving the estimating equation

$$\sum_{i=1}^{n} \mathbf{G}_i^{\mathrm{id}}(\mathbf{d}_i, \boldsymbol{\omega}) = \sum_{i=1}^{n} (\mathbf{d}_i - \bar{\mathbf{d}}_i)\,(\mathbf{Y}_i - \mathbf{L}_i\boldsymbol{\omega}) = \mathbf{0} \tag{3.3}$$

where $\mathbf{d}_i = \mathbf{d}\,(i, \mathbf{L}_i, S_i)$ (which depends on the cluster index $i$ only through its dimension) is an arbitrary $q \times n_i$ matrix function of $(\mathbf{L}_i, S_i)$ with $\mathbf{L}_i = (\mathbf{L}_{i1}, ..., \mathbf{L}_{in_i})'$. Further, $q$ is the number of estimable parameters in $\boldsymbol{\omega}$ and $\bar{\mathbf{d}}_i$ is a $q \times n_i$ matrix with the sample averages $\sum_{j=1}^{n_i} d_{ikj}/n_i$ in the $k$th row, where $d_{ikj}$ is the element in the $k$th row and $j$th column of $\mathbf{d}_i$. Subtracting the cluster mean $\bar{\mathbf{d}}_i$ of $\mathbf{d}_i$ ensures that for arbitrary $\mathbf{d}\,(i, \mathbf{L}_i, S_i)$, the estimating functions involve within-cluster comparisons (i.e. differences between residuals $Y_{ij} - \alpha - X_{ij}\beta - V_{ij}\gamma - S_i\delta - b_i$ within cluster $i$, so that the cluster-level part $-\alpha - S_i\delta - b_i$ disappears) and therefore yield protection against the presence of unmeasured cluster-level confounders. For example, with follow-up data obeying model

$$E(Y_{it}|\mathbf{X}_i, b_i) = \alpha_1 + t\alpha_2 + X_{it}\beta + X_{i,t-1}\gamma + b_i \tag{3.4}$$

where $t = 1, ..., n_i$ and $X_{i0} = 0$, solving (3.3) with $\mathbf{d}_i = (\mathbf{t}, \mathbf{X}_i, \mathbf{X}_{i,-1})'$, $\mathbf{t} = (1, ..., n_i)'$ and $\mathbf{X}_{i,-1} = (X_{i0}, ..., X_{i,n_i-1})'$, is equivalent to solving generalized estimating equations (GEE) with independence working correlation under model

$$E(Y_{it} - \bar{Y}_i|\mathbf{X}_i) = (t - \bar{t})\alpha_2 + (X_{it} - \bar{X}_i)\beta + (X_{i,t-1} - \bar{X}_{i,-1})\gamma \tag{3.5}$$

where $\bar{X}_{i,-1} = \sum_{t=1}^{n_i} X_{i,t-1}/n_i$. This model is obtained by subtracting the cluster average from the left and righthand side of (3.4). While the usefulness of methods which regress changes in response on changes in covariates has long been realized (Louis et al., 1986), Theorem 3, Part 1b, shows that the resulting estimators are essentially the only CAN estimators for $\boldsymbol{\omega}$ under model (3.2). Specifically, the efficient estimator for $\boldsymbol{\omega}$ under model (3.2), as given in Theorem 4, can be obtained from a regression of changes.

Suppose now that $h(.)$ is the exponential link. Let $\widetilde{\mathbf{Y}}_i(\boldsymbol{\omega})$ be the $n_i \times 1$ vector with $j$th component $\widetilde{Y}_{ij}(\boldsymbol{\omega}) = Y_{ij}\exp(-\mathbf{L}_{ij}\boldsymbol{\omega})$. Part 2 of Theorem 3 then shows that, up to asymptotic equivalence, all CAN estimators for $\boldsymbol{\omega}$ under the observed data model (3.2) can be obtained by solving the estimating equation

$$\sum_{i=1}^{n} \mathbf{G}_i^{\log}(\mathbf{d}_i, \boldsymbol{\omega}) = \sum_{i=1}^{n} (\mathbf{d}_i - \bar{\mathbf{d}}_i)\widetilde{\mathbf{Y}}_i(\boldsymbol{\omega}) = \mathbf{0} \tag{3.6}$$

where $\mathbf{d}_i = \mathbf{d}(i, \mathbf{L}_i, S_i)$ is defined as before. This equation expresses that, after having removed the cluster-varying part $\exp(\mathbf{L}_{ij}\boldsymbol{\omega})$ from the outcome, it should retain no cluster-varying components and thus $\widetilde{\mathbf{Y}}_i(\boldsymbol{\omega})$ should have zero covariance with $\mathbf{d}_i - \bar{\mathbf{d}}_i$. Specifically, with cross-sectional data for pairs (i.e. $n_i = 2, \forall i$) obeying the model $E(Y_{ij}|\mathbf{X}_i) = \exp(\alpha + X_{ij}\beta)$ and $\mathbf{d}_i = (X_{i1}, X_{i2})$, the estimating equation in (3.6) can be written as

$$\sum_{i=1}^{n} (X_{i1} - \bar{X}_i)Y_{i1}\exp(-X_{i1}\beta) + (X_{i2} - \bar{X}_i)Y_{i2}\exp(-X_{i2}\beta) = 0$$

Part 3 of Theorem 3 shows that, owing to non-collapsibility of the odds ratio, no $\sqrt{n}$-consistent estimators exist for $\boldsymbol{\omega}$ under the observed data model (3.2) when $h(.)$ is the inverse logit link. Informally, this is because no reduced model of within-cluster differences, such as (3.5), exists for the logit link. It follows that it is not possible to obtain estimators for the within-cluster effects $\boldsymbol{\omega}$ that are protected against unmeasured cluster-level confounders and converge at the usual rate under this model, unless one is willing to make additional assumptions. For instance, under the additional assumption that outcomes within clusters are independent (given measured covariates), it is well known that conditional likelihood estimation yields $\sqrt{n}$-consistent estimators for $\boldsymbol{\omega}$ under the resulting, more restrictive model. A more general development for the logistic model that allows general correlation structures is beyond the scope of this work.

**Theorem 3.** *Suppose that the regularity conditions in Appendix 3.A1 hold. Then,*

1. *(a) under the observed data model (3.2) with $h(.)$ equalling the identity link, the estimating equations (3.3) have a solution $\widehat{\boldsymbol{\omega}}(\mathbf{d})$ such that*

$\sqrt{n}(\widehat{\boldsymbol{\omega}}(\mathbf{d}) - \boldsymbol{\omega}) \xrightarrow{d} N(0, \Sigma(\mathbf{d}))$. $\Sigma(\mathbf{d})$ *can be consistently estimated with*

$$E_n\left\{\frac{\partial \mathbf{G}_i^{id}(\mathbf{d}_i, \widehat{\boldsymbol{\omega}}(\mathbf{d}))}{\partial \boldsymbol{\omega}}\right\}^{-1} E_n\left\{\mathbf{G}_i^{id}(\mathbf{d}_i, \widehat{\boldsymbol{\omega}}(\mathbf{d}))^{\otimes 2}\right\} E_n\left\{\frac{\partial \mathbf{G}_i^{id}(\mathbf{d}_i, \widehat{\boldsymbol{\omega}}(\mathbf{d}))}{\partial \boldsymbol{\omega}'}\right\}^{-1} \tag{3.7}$$

*where for any random variable* $\mathbf{W}$, $E_n(\mathbf{W}) = \sum_{i=1}^{n} W_i/n$ *and* $\mathbf{W}^{\otimes 2} = \mathbf{W}\mathbf{W}'$.
*(b) if* $\widehat{\widehat{\boldsymbol{\omega}}}$ *is a CAN estimator of* $\boldsymbol{\omega}$ *under the observed data model (3.2) with* $h(.)$ *the identity link, then there exists a* $q \times n_i$ *vector function* $\mathbf{d}(i, \mathbf{L}_i, S_i)$ *such that* $\sqrt{n}\left(\widehat{\boldsymbol{\omega}}(\mathbf{d}) - \widehat{\widehat{\boldsymbol{\omega}}}\right)$ *converges to 0 in probability.*

2. *part 1 of Theorem 3 holds with* $h(.)$ *replaced by the inverse log link, estimating equations (3.3) replaced by (3.6) and* $\mathbf{G}_i^{id}(\mathbf{d}_i, \widehat{\boldsymbol{\omega}}(\mathbf{d}))$ *by* $\mathbf{G}_i^{log}(\mathbf{d}_i, \widehat{\boldsymbol{\omega}}(\mathbf{d}))$.

3. *no* $\sqrt{n}$*-consistent estimators exist for* $\boldsymbol{\omega}$ *under the observed data model (3.2) with* $h(.)$ *equalling the inverse logit link.*

Note that because the true association of measured cluster-level confounders $S_i$ with outcome in model (3.2) can be considered to be part of $b_i$, the results of Theorem 3 continue to hold when that association is misspecified.

Throughout we will refer to equations (3.3) and (3.6) as conditional generalized estimating equations (CGEE). These represent a subset of the generalized estimating equations corresponding to model (3.2) with $b_i$ empty (i.e. set to zero). Consequently, estimators for $\boldsymbol{\omega}$ in model (3.2) obtained by solving CGEE are never more efficient than those obtained by ordinary GEE when $b_i$ is independent of within-cluster covariates. However, unlike GEE-estimators, they are guaranteed to be asymptotically unbiased even when $b_i$ is associated with within-cluster covariates. In our opinion, the hope to control bias trumps efficiency concerns and it is because of this that we recommend using CGEE instead of ordinary GEE when estimating the effects of within-cluster exposures in the possible presence of important

unmeasured cluster-level confounders and/or with concern for misspecification of the association of cluster-level confounders with outcome. Because one may lose efficiency, it becomes increasingly important to identify and implement the efficient estimating function (i.e. the efficient score) within our class. In Theorem 4, we show that when $h(.)$ is the identity link, the efficient score for $\boldsymbol{\omega}$ under our model equals $\mathbf{G}_i^{\mathrm{id}}(\mathbf{d}_{eff}, \boldsymbol{\omega})$ with

$$\mathbf{d}_{eff}(i, \mathbf{L}_i, S_i) = \mathbf{L}_i' \mathrm{Var}^{-1}(\mathbf{Y}_i | \mathbf{L}_i, C_i) \tag{3.8}$$

when the within-cluster variance $\mathrm{Var}(\mathbf{Y}_i|\mathbf{L}_i, C_i)$ (with $C_i$ denoting the index of cluster $i$) is constant. When $h(.)$ is the inverse log link, the efficient score equals $\mathbf{G}_i^{\log}(\mathbf{d}_{eff}, \boldsymbol{\omega})$ with

$$\mathbf{d}_{eff}(i, \mathbf{L}_i, S_i) = \mathbf{L}_i' \mathrm{Var}^{-1}(\widetilde{\mathbf{Y}}_i(\boldsymbol{\omega}) | \mathbf{L}_i, C_i) \overline{\widetilde{\mathbf{Y}}_i(\boldsymbol{\omega})} \tag{3.9}$$

and $\overline{\widetilde{\mathbf{Y}}_i(\boldsymbol{\omega})} = \sum_{j=1}^{n_i} \widetilde{Y}_{ij}(\boldsymbol{\omega})/n_i$, when the within-cluster variance $\mathrm{Var}(\widetilde{\mathbf{Y}}_i(\boldsymbol{\omega})|\mathbf{L}_i, C_i)$ is constant. When outcomes are independent within each cluster, then $\mathrm{Var}(\mathbf{Y}_i|\mathbf{L}_i, C_i)$ is the $n_i \times n_i$ identity matrix in which case the efficient choice for $\mathbf{d}_i$ simplifies to $\mathbf{L}_i'$ for the identity link. This choice is also efficient for the inverse log link when, in addition, the within-cluster mean and variance of the outcome are the same (as would be the case for Poisson data). For most practical purposes the latter choices will provide good approximations of the efficient score since the within-cluster correlation is typically weak as compared to the correlation induced by between-cluster components.

**Theorem 4.** *Estimating function* $\mathbf{G}_i^{id}(\mathbf{d}_{eff}, \boldsymbol{\omega})$ *with* $\mathbf{d}_{eff}$ *as defined in (3.8) when* $h(.)$ *is the identity link and estimating function* $\mathbf{G}_i^{log}(\mathbf{d}_{eff}, \boldsymbol{\omega})$ *with* $\mathbf{d}_{eff}$ *as defined in (3.9) when* $h(.)$ *is the inverse log link, is the efficient score for* $\boldsymbol{\omega}$ *under the observed data model (3.2) in the sense that for any* $\mathbf{d}$, $Var\{\widehat{\boldsymbol{\omega}}(\mathbf{d})\} \geq Var\{\widehat{\boldsymbol{\omega}}(\mathbf{d}_{eff})\}$, *provided that the within-cluster variance* $Var(\mathbf{Y}_i|\mathbf{L}_i, C_i)$ *or* $Var(\widetilde{\mathbf{Y}}_i(\boldsymbol{\omega})|\mathbf{L}_i, C_i)$, *respectively, is constant.*

In Appendix 3.A4, we deduce from (3.8) that, under the identity link, the efficient score can be easily obtained via a regression of changes using generalized estimating equations with working covariance equal to the covariance of the changes $\mathbf{Y}_i - \overline{\mathbf{Y}}_i$. Because the latter covariance is more

difficult to specify than the covariance of the original outcome, we recommend however to work on the original scale, as suggested by expression (3.8). To estimate the within-cluster variance matrix in this expression, we propose to fit a random effects model for the outcome and to use the estimated residual covariance matrix (i.e. the outcome covariance, given the random effects) as an estimate for $\text{Var}(\mathbf{Y}_i|\mathbf{L}_i, C_i)$. When $h(.)$ is the inverse log link, we proceed in two steps because the efficient choice for $\mathbf{d}_i$ then involves the unknown parameter. First we obtain an inefficient estimate $\hat{\boldsymbol{\omega}}$ for $\boldsymbol{\omega}$ by choosing $\mathbf{d}_i = \mathbf{L}_i'$ in (3.6). Next, we fit a random effects model for the transformed outcome $\widetilde{\mathbf{Y}}_i(\hat{\boldsymbol{\omega}})$ and use the estimated residual covariance matrix as an estimate for $\text{Var}(\widetilde{\mathbf{Y}}_i(\boldsymbol{\omega})|\mathbf{L}_i, C_i)$.

## 3.4 Separating within- from between-cluster exposure effects

Neuhaus and Kalbfleisch (1998) propose to estimate the effect of within-cluster exposures in the presence of unmeasured cluster-level confounders by separating exposure effects into within- and between-cluster components, as in model (3.1). Several studies have shown by example that for $h$ the identity or inverse logit link, ordinary mixed-effects estimation of the within-cluster exposure effect $\beta_w$ yields an estimate that is nearly identical to the one obtained via conditional likelihood estimation, and theoretical justifications have been reported (Neuhaus and McCulloch, 2006). The simplicity of this approach has made it a popular alternative to conditional likelihood methods (Ten Have et al., 2004), even though its validity and efficiency has not been formally studied.

In Appendix 3.A4 we consider GEE-estimators obtained by fitting model

$$E(Y_{ij}|\mathbf{X}_i, \mathbf{V}_i) = h\{(X_{ij} - \bar{X}_i)\beta + (V_{ij} - \bar{V}_i)\gamma\}, \tag{3.10}$$

which replaces each within-cluster exposure by its deviation from the cluster mean. We show that, when $h(.)$ is the identity link, all these estimators belong to our class of CGEE-estimators under model (3.2) (regardless of

whether one adds an intercept and further adjusts for the cluster mean of the within-cluster exposures), provided that the chosen working model for the outcome covariance does not allow dependence on cluster-varying covariates. Under this restriction, it thus follows that the NK-approach is valid for the identity link. This approach also yields efficient estimators under the data-generating model (3.2) when the working covariance equals the true outcome covariance, suggesting that CGEE-estimators are not more efficient. However, when $h(.)$ is the exponential link, we show in Appendix 3.A4 that estimators obtained by fitting model (3.10) may be inconsistent for the effect of within-cluster exposures in model (3.2). Informally, this is because no reduced conditional mean model of within-cluster differences can be derived due to the nonlinearity of the log link. We therefore recommend to solve CGEE in that setting. For the logit link, it follows from Part 3 of Theorem 3 that no $\sqrt{n}$-consistent estimators can be found for the effect of within-cluster exposures in model (3.2). Specifically, fitting model (3.10) with $h(.)$ the inverse logit link will yield no $\sqrt{n}$-consistent estimators under model (3.2) (unless possibly under additional assumptions on the outcome covariance).

Confusion has been raised regarding the interpretation of the parameters in models such as (3.1) because these models involve the target exposure only through its deviation from cluster mean (Begg and Parides, 2003). Specifically, $\beta_w$ in model (3.1) represents the expected change in outcome when increasing $X_{ij}$ with one unit, while holding the cluster mean fixed, which is difficult to interpret because the cluster mean itself involves $X_{ij}$. The discussion in the previous paragraph shows that for the identity link, estimates for the within-cluster effects (i.e. the effects of $X_{ij} - \bar{X}_i$) in these models can be interpreted as within-cluster effects in standard conditional mean models. Our focus on such standard models in this article is additionally motivated by the fact that models which involve within- and between-cluster exposure effects cannot be viewed as data-generating models and are therefore biologically/causally difficult to interpret. For instance, conditioning on $\bar{X}_i$ in longitudinal studies is tantamount to letting the future determine the present because $\bar{X}_i$ involves measurements up to the study end. Likewise models that condition on $\bar{X}_i$ in cross-sectional

clustered sampling designs cannot be viewed as data-generating models in settings where the outcome cannot be affected by other subjects' covariates. Further, even if other subjects' covariates affect outcome, the meaning of the model parameters in (3.1) remains unclear. It can only be deduced by reparameterizing the model as

$$\mathrm{E}(Y_{ij}|b_i, \mathbf{X}_i) = h\left\{\beta_0 + b_i + \left(\beta_w + \frac{\beta_b - \beta_w}{n_i}\right) X_{ij} + \frac{\beta_b - \beta_w}{n_i} \sum_{k \neq j} X_{ik}\right\} \tag{3.11}$$

that $\beta_w$ does not represent the effect of one's own exposure, but the difference in effect between one's own exposure and the exposure of other cluster members. Because models that involve within- and between-cluster effects cannot be viewed as data-generating models, postulating them can be especially difficult in complex settings, such as family-based studies where parents' exposure affects the offspring's outcome differently than siblings' exposure, and in longitudinal studies where exposure effects persist, but weaken over time.

Besides offering computational simplicity, the NK-approach allows estimation of the effect of cluster-level exposures. Such effects must however be cautiously interpreted. First, in cross-sectional study designs where outcome cannot be affected by other cluster members' exposure, it follows from (3.11) that $\beta_w$ should equal $\beta_b$. Any deviation of $\beta_b$ from $\beta_w$ is therefore indicative of the presence of (unmeasured) cluster-level confounders. Likewise, any such deviation in longitudinal studies reflects bias due to confounding since the model would otherwise allow future exposures to affect present outcome. Second, also the effects of cluster-level variables (e.g. $S_i$) on outcome will typically carry no causal meaning. This is not only because of similar confounding bias, but also because within-cluster exposures will often be intermediate on the causal path from cluster-level exposures to outcome and adjustment for such intermediate variables may induce selection bias (Hernan et al., 2004). If the association of cluster-level exposures were nevertheless of interest, note that they can be estimated using standard GEE estimation under model (3.2) with $b_i$ set to zero and the effects of within-cluster exposures replaced by their CGEE-estimates.

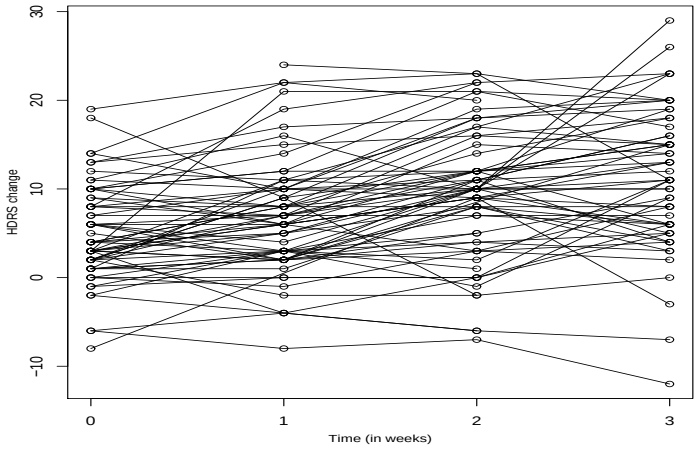


Figure 3.1: Profile plot of depression change versus time.

## 3.5 Data analysis

In this section, we analyze the data described in Section 3.2 to examine the clinical effect of IMI on the expected change in depression. All analyses were conducted in R (version 2.3.1). The profile plot in Figure 3.1 shows a clear indication of a random intercept, but little or no evidence of a random time evolution. Thus, we start from the conditional mean model

$$E(Y_{ij}|\mathbf{X}_i, b_i) = \alpha + X_{ij}\beta + t_{ij}\gamma + b_i, \tag{3.12}$$

where $Y_{ij}$ is the change in HDRS, $X_{ij}$ the log concentration ratio of DMI and IMI, and $t_{ij}$ the time point (in weeks) at which the $j$th measurement was recorded for patient $i$. Throughout we will assume that IMI absorption is not affected by the previous depression status (i.e. change in HDRS) and that there are no remaining time-varying confounders for the association between IMI absorption and change in HDRS.

We first fitted model (3.12) using GEE and efficient CGEE, assuming no within-cluster correlation. Next, in line with the NK-approach, we fitted

model

$$E(Y_{ij}|\mathbf{X}_i, b_i) = \alpha + \beta_w(X_{ij} - \overline{X}_i) + \beta_b\overline{X}_i + \gamma t_{ij} + b_i, \tag{3.13}$$

using GEE with compound symmetry working covariance. The 3 results are summarized in the first 3 columns of Table 3.1 and are labeled GEE, CGEE and NK, respectively. The CGEE-analysis shows that, at a fixed time point, the effect of a unit increase in log concentration ratio (i.e. better absorption of IMI) is to increase (i.e. improve) the expected change in HDRS with 0.57 (95% confidence interval (CI) $[-1.68, 2.82]$). As predicted by the theory, it yields results similar to those obtained from model (3.13). In contrast, the GEE-analysis which ignores the possibility of cluster-level confounding, yields a doubled effect size which is almost significantly different from zero (1.18 with 95% CI $[-0.05, 2.41]$). The difference between these results is suggestive of the possible presence of patient-specific confounders which may compromise the GEE analysis. The advantage of obtaining more robust estimates with CGEE comes, however, with a price of reduced efficiency. In the next section, we will investigate this via simulation studies.

We refitted model (3.12) using CGEE allowing for a residual homogeneous autoregressive covariance structure besides the exchangeable correlation implied by a random intercept, and similarly for model (3.13). This further decreased the clinical effect of IMI to 0.50 with CGEE, but gave no change in precision. This is not surprising because the residual correlation was estimated to equal merely 0.18 using CGEE. We further observed an expected decrease of 2.05 (95% CI $[1.29, 2.81]$) in HDRS per week using CGEE, suggesting that the average depression status improves over time. There was no indication that the clinical effect of IMI changes over time (p-value 0.12 using CGEE).

To investigate whether the exposure effect differs between patients with/without a diagnosis of endogenous depression at baseline ($= S_i$), we added a main effect and interaction of diagnosis with exposure to all models. Following Section 3.4, this can be done by subtracting the cluster average from the cluster-varying exposures $X_{ij}$ and $X_{ij}S_i$:

$$E(Y_{ij}|\mathbf{X}_i, S_i) = \alpha + \beta_w(X_{ij} - \overline{X}_i) + \beta_b\overline{X}_i + \gamma t_{ij} + \delta S_i + \beta(X_{ij}S_i - \overline{X_iS_i}).$$

This gives similar results as using efficient CGEE (i.e. (3.3) with $\mathbf{L}_i = (\mathbf{X}_i, \mathbf{t}_i, \mathbf{X}_i S_i)$ and $\mathbf{d}_i = \mathbf{d}_{eff}$). Table 3.2 reveals a significant IMI effect among patients with non-endogenous depression using GEE, but not when the possible presence of cluster-level confounders is taken into account. None of the 3 methods find a significant interaction effect, but the estimated value obtained with GEE is more than three times smaller than with CGEE.

| | Exchangeable correlation | | | Random intercept and autoregressive residual correlation | | |
|---|---|---|---|---|---|---|
| Working model | CGEE | GEE | NK | CGEE | GEE | NK |
| IMI (SE) | 0.57 (1.15) | 1.18 (0.63) | 0.57 (1.15) | 0.50 (1.15) | 1.18 (0.63) | 0.48 (1.15) |
| p-value | 0.62 | 0.061 | 0.62 | 0.66 | 0.062 | 0.68 |
| time (SE) | 2.01 (0.37) | 2.01 (0.38) | 2.03 (0.37) | 2.05 (0.39) | 2.13 (0.42) | 2.14 (0.41) |
| p-value | <0.0001 | <0.0001 | <0.0001 | <0.0001 | <0.0001 | <0.0001 |

Table 3.1: *Estimates, standard errors (SE) and p-values for within-cluster effects in model (3.12) using CGEE, using GEE, and in model (3.13) using the NK-approach (NK).*

| | Exchangeable correlation | | | Random intercept and autoregressive residual correlation | | |
|---|---|---|---|---|---|---|
| Working model | CGEE | GEE | NK | CGEE | GEE | NK |
| IMI (SE) | 2.14 (1.11) | 1.96 (0.82) | 2.14 (1.10) | 2.04 (1.09) | 1.68(0.81) | 1.85 (1.06) |
| p-value | 0.050 | 0.017 | 0.053 | 0.060 | 0.040 | 0.083 |
| time (SE) | 2.14 (0.35) | 2.06 (0.36) | 2.16 (0.35) | 2.18 (0.37) | 2.16 (0.40) | 2.25 (0.38) |
| p-value | <0.0001 | <0.0001 | <0.0001 | <0.0001 | <0.0001 | <0.0001 |
| IMI × diagn (SE) | -3.90 (2.33) | -1.60 (1.32) | -3.91 (2.32) | -3.68 (2.36) | -0.94 (1.27) | -3.19 (2.36) |
| p-value | 0.090 | 0.22 | 0.092 | 0.12 | 0.46 | 0.18 |

Table 3.2: *Estimates, standard errors (SE) and p-values for within-cluster effects in model (3.12) using CGEE, using GEE, and in model (3.13) using the NK-approach (NK), where all models are augmented with a main effect of diagnosis and its interaction with IMI.*

## 3.6 Simulation study

To investigate the bias-variance trade-off of using CGEE versus GEE, we generated 500 data sets using the following models:

$$Y_{ij} = h(\alpha_Y + X_{ij}\beta_Y + S_i\delta_Y + b_{iY} + \epsilon_{ijY}) \quad (3.14)$$

$$X_{ij} = \alpha_X + S_i\delta_X + \eta b_{iX} + \epsilon_{ijX} \quad (3.15)$$

where $h$ is the identity link or the inverse log link, $S_i$ is normally distributed with mean $\mu_S$ and standard deviation $\sigma_S$, $b_{iX}$ and $b_{iY}$ are independent normally distributed mean zero random intercepts with standard deviations $\sigma_{bX}$ and $\sigma_{bY}$, and $\epsilon_{ijX}$ and $\epsilon_{ijY}$ are independent normally distributed mean zero residual errors with standard deviations $\sigma_{wX}$ and $\sigma_{wY}$. We conducted 6 simulation studies, each time with and without the presence of unmeasured cluster-level confounders. We define $\theta^1_{id} = (\alpha_Y, \beta_Y, \delta_Y, \sigma_{bY}, \sigma_{wY}, \alpha_X, \delta_X, \eta, \sigma_{wX}, \sigma_{bX}, \mu_S, \sigma_S)$, $\theta^2_{id} = (\delta_Y, \alpha_X, \delta_X, \eta, \sigma_{bX})$, $\theta^1_{log} = (\alpha_Y, \delta_Y, \sigma_{bY}, \alpha_X, \delta_X, \eta, \mu_S, \sigma_S)$ and $\theta^2_{log} = (\delta_Y, \delta_X, \eta, \sigma_{bX})$. In simulation 1, we used 400 clusters of size 5 with identity link, uncorrelated residuals $\epsilon_{ijY}$ and $\theta^1_{id} = (1, 2, 1.2, 2, 1, 2, -1, 0, 0.5, 0, 2, 2)$ (and $\theta^2_{id} = (0, 3, 0, 1, 1.5)$ in the analysis without confounder). In simulations 2 and 3, we used 50 clusters of size 4 with identity link, correlated residuals $\epsilon_{ijY}$ (with exponentially decaying correlation of 0.6 between successive time points) and $\theta^1_{id} = (3, 1, 2, 1, 5, 1, 3, 0, 1, 0, 4, 3)$ (and $\theta^2_{id} = (0, 1, 0, 1, 2)$ in the analysis without confounder). In simulations 4 and 5, we used 500 clusters of size 3 with inverse log link, independent Poisson distributed outcomes conditional on $b_{iY}$, and $\theta^1_{log} = (0, -0.5, 0.5, 0, -0.5, 0, 0, 1)$ (and $\theta^2_{log} = (0, 0, 0.5, 2)$ in the analysis without confounder). In simulation 6, we used 400 clusters of size 5 with inverse log link, independent Poisson distributed outcomes conditional on $b_{iY}$, and $\theta^1_{log} = (0, 0.6, 0.5, 2, -0.3, 0, 0, 2)$ (and $\theta^2_{log} = (0, 0, 0.5, 2)$ in the analysis without confounder). For $(\beta_Y, \sigma_{wX})$ we chose values $(1.2, 0.7)$, $(0.6, 2)$ and $(0.4, 0.5)$ for simulation 4, 5 and 6, respectively.

The simulated data were analyzed using the following 3 approaches, each time using the same working covariance model: (a) CGEE with compound symmetry working covariance for simulations 1, 3 and 6, with independence working covariance for simulation 4 and 5, and using a random

intercept model with autoregressive residual covariance for simulation 2; (b) GEE, ignoring the presence of the cluster-level confounder $S_i$ (i.e. setting $\delta_Y = 0$); (c) NK, by fitting model $E(Y_{ij}|\mathbf{X}_i) = \alpha + \beta_w(X_{ij} - \overline{X}_i) + \beta_b\overline{X}_i$ using GEE (for reasons of comparability with the other estimators).

The results are summarized in Tables 3.3 and 3.4. As expected, GEE gives seriously biased results when a cluster-level confounder is ignored, while CGEE yields unbiased estimates. When the model is linear and there are no unmeasured confounders, then, as expected, the GEE estimator is more precise than the CGEE estimator with relative efficiencies of 1.33, 1.96 and 2.04 for simulation 1, 2 and 3, respectively. Although substantial, this may be affordable, given the huge loss of coverage of the GEE estimator in the presence of unmeasured confounding. We further observe that, as predicted by the theory, the NK approach gives similar results as CGEE for the identity link. For the inverse log link, the NK approach performs well when the effect size $\beta_Y$ and the within-cluster standard deviation in $X$ are small, as in simulation 6, but has a worse performance than the CGEE estimator otherwise in terms of bias, precision (relative efficiencies of 2.64 and 2.15 in simulation 4 and 5, respectively) and coverage. Surprisingly, the CGEE estimator has greater precision than the GEE estimator in simulations 4 and 5 in the absence of unmeasured confounding. This is a finite sample problem caused by the fact that the GEE (and NK) estimator require estimation of the covariance structure, unlike the CGEE estimator. Indeed, the precision advantage of the CGEE estimator disappeared in additional simulation studies (not displayed) with larger numbers of clusters or in the absence of correlation (so that the independence model holds).

| | Bias | | | Empirical SE | | | Coverage | | |
|---|---|---|---|---|---|---|---|---|---|
| Simulation | CGEE | GEE | NK | CGEE | GEE | NK | CGEE | GEE | NK |
| 1 | 0.0014 | 0.48** | 0.0014 | 0.053 | 0.055 | 0.053 | 0.93 | 0.00 | 0.93 |
| 2 | 0.0080 | 0.63** | 0.0050 | 0.28 | 0.067 | 0.27 | 0.93 | 0.00 | 0.93 |
| 3 | -0.0070 | 0.64** | -0.0070 | 0.30 | 0.068 | 0.30 | 0.95 | 0.00 | 0.95 |
| 4 | -0.00099 | 0.29** | -0.022** | 0.032 | 0.052 | 0.052 | 0.95 | 0.00067 | 0.93 |
| 5 | -0.000042 | 0.034** | -0.020** | 0.015 | 0.022 | 0.022 | 0.95 | 0.75 | 0.89 |
| 6 | 0.0011 | -0.12** | 0.0022 | 0.025 | 0.020 | 0.026 | 0.94 | 0.00 | 0.94 |

Table 3.3: *Simulation results for $\beta$ (with unmeasured cluster-level confounder), ** = p-value (for $H_0$: bias=0) <0.0001*

| | Bias | | | Empirical SE | | | Coverage | | |
|---|---|---|---|---|---|---|---|---|---|
| Simulation | CGEE | GEE | NK | CGEE | GEE | NK | CGEE | GEE | NK |
| 1 | 0.0014 | 0.0022 | 0.0014 | 0.053 | 0.046 | 0.053 | 0.93 | 0.97 | 0.93 |
| 2 | 0.0080 | -0.0030 | 0.0047 | 0.28 | 0.20 | 0.27 | 0.93 | 0.93 | 0.93 |
| 3 | -0.0070 | -0.0026 | -0.0070 | 0.30 | 0.21 | 0.30 | 0.95 | 0.92 | 0.95 |
| 4 | -0.00021 | -0.0024 | 0.0021 | 0.030 | 0.051 | 0.051 | 0.95 | 0.87 | 0.93 |
| 5 | 0.00048 | 0.00030 | 0.00061 | 0.014 | 0.019 | 0.022 | 0.94 | 0.92 | 0.93 |
| 6 | -0.00052 | -0.0025$^\star$ | -0.0018$^\star$ | 0.018 | 0.022 | 0.019 | 0.96 | 0.92 | 0.96 |

Table 3.4: *Simulation results for $\beta$ (without unmeasured cluster-level confounder), $^\star$ = p-value (for $H_0$: bias=0) <0.05*

## 3.7 Discussion

Conditional generalized estimating equations provide a general framework for protecting estimation of within-cluster exposure effects in the analysis of clustered and longitudinal data, against unmeasured cluster-level confounding factors and against misspecification of the effects of measured cluster-level variables. They are obtained by using G-estimation (Robins, Mark and Newey, 1992) to avoid modelling the association of cluster-level confounders with outcome, while instead modelling their association with exposure nonparametrically in terms of cluster-level averages. Paired with the theory of G-estimation, we find that such protection against unmeasured cluster-level confounding factors is possible under the identity and log link, but not under the logistic link without additional assumptions (Vansteelandt and Goetghebeur, 2003).

By construction, other common types of estimators that offer protection against unmeasured cluster-level confounders can be viewed as estimators within our class, e.g. conditional likelihood estimators (Diggle, Liang and Zeger, 1994; Verbeke, Spiessens and Lesaffre, 2001), difference-in-difference estimators (Abadie, 2005),... Our approach extends these to general covariance structures and nonlinear link functions under weaker assumptions. It avoids separation of exposures into within- and between-cluster components (Neuhaus and Kalbfleisch, 1998) because (a) models which involve such components cannot be viewed as data-generating models; and (b) such approaches may be inconsistent and inefficient under nonlinear link functions. Nonetheless, we show the latter approaches to be very useful and attractive because they offer computationally convenient estimators, they are valid under the identity link and approximately valid, but inefficient, under the log link. R-programs for solving CGEE can be obtained upon request. Future work will concentrate on whether and how results can be extended to longitudinal studies with measured time-varying confounders.

# Appendix 3.A1: Assumptions

Throughout the article, we make the following assumptions which are required to ensure that $\beta$ in model (3.2) represents the causal effect of exposure $X_{ij}$ on outcome $Y_{ij}$. When the study is cross-sectional and each cluster represents observations obtained from different but related subjects, we will assume that, given $(X_{ij}, V_{ij}, S_i, b_i)$, the outcome for subject $j$ in cluster $i$ has no residual dependence on within-cluster variables $(X_{ik}, V_{ik}), k \neq j$ of other subjects in that cluster. When the study is longitudinal, we will assume that the process $(X_{ij}, V_{ij})$ is ancillary (Robins, 1999a), meaning that $Y_{ij}$ is conditionally independent of future covariates after adjustment for the covariate history. This assumption is valid whenever covariates $(X, V)$ are not affected by previous outcomes and there are no remaining time-varying confounders for the association between $(X, V)$ and outcome. In both study designs, when the goal is to infer the effect of exposure on outcome, we will assume that it is sufficient to adjust for measured within-cluster confounders $V_{ij}$, i.e. that all within-cluster confounders have been measured and correctly adjusted for. These assumptions are implicit in the NK-approach.

# Appendix 3.A2: proof of Theorem 3

For notational convenience, we will ignore boldface notation to distinguish vectors and scalars in this appendix. Suppose first that $b$ is completely observed. The nuisance tangent space (Bickel et al., 1993) $\Lambda^F = \overline{\Lambda_1^F + \Lambda_2^F + \Lambda_3^F}$ for the full data model $\mathcal{A}^F$ defined by the restrictions of model

$$E(Y|X,V,S,b) = h\left\{\alpha + \beta X + \gamma V + \delta S + g(b;\eta_3)\right\}$$

where $g(.)$ is an unknown function, is then the closed linear span of the union of the tangent sets (Bickel et al., 1993) $\Lambda_1^F$, $\Lambda_2^F$, $\Lambda_3^F$ corresponding to the parameters in any regular parametric submodel for $f(Y|X,V,S,b)$, $f(X,V,S,b)$ and $g(b)$, respectively.

It follows from Bickel et al. (1993) that the orthocomplement of the tangent space $\Lambda_1^F + \Lambda_2^F$ is $\left(\Lambda_1^F + \Lambda_2^F\right)^{\perp} = \{D\,\epsilon(\omega,\nu,\eta_3)\ :\ D = d\,(X,V,S,b)$ arbitrary$\}$. To derive the restrictions on the elements of the set $\Lambda_3^F$, consider arbitrary parametric submodels $g(b;\eta_3)$ containing the true model. Denote $\epsilon(\omega,\nu,\eta_3) = Y - h\left\{\alpha + \beta X + \gamma V + \ \delta S + g(b;\eta_3)\right\}$ where $\nu = (\alpha,\delta)'$. From the expressions for the full data scores w.r.t. $\eta_3$, we find that

$$\Lambda_3^F = \{A \equiv a(Y|X,V,S,b)\ :\ E(A|X,V,S,b) = 0;$$
$$E\left\{\epsilon(\omega,\nu,\eta_3)A|X,V,S,b\right\} = h'\{\alpha + \beta X + \gamma V + \delta S + g(b;\eta_3)\}\frac{\partial g(b;\eta_3)}{\partial \eta_3}\Big\}$$

where $h'(x) = \frac{\partial h(x)}{\partial x}$. To derive the set of all influence functions for $\omega$ indexing $\mathcal{A}^F$, we use the fact that $\left(\Lambda_1^F + \Lambda_2^F + \Lambda_3^F\right)^{\perp} = \left(\Lambda_1^F + \Lambda_2^F\right)^{\perp} \cap\ \Lambda_3^{F\perp}$

and hence identify those functions $D = d(X, V, S, b)$ satisfying for each $A \equiv a(Y|X, V, S, b) \in \Lambda_3^F$:

$$\begin{aligned} 0 &= E\{D\,\epsilon(\omega, \nu, \eta_3)A\} \\ &= E\left[D\,h'\{\alpha + \beta X + \gamma V + \delta S + g(b; \eta_3)\}\frac{\partial g(b; \eta_3)}{\partial \eta_3}\right] \end{aligned} \tag{3.16}$$

Consider first model $\mathcal{A}^F$ with $h(.)$ equalling the identity link. Then the above equality (3.16) is satisfied if and only if $E(D|b) = 0$. Hence, for $D = d(X, V, S, b)$

$$\left(\Lambda_1^F + \Lambda_2^F + \Lambda_3^F\right)^{\perp} = \{D\,\epsilon(\omega, \nu, \eta_3) : D \text{ arbitrary satisfying} E(D\,|b) = 0\}$$

Consider now the observed data model $\mathcal{A}$, implied by $\mathcal{A}^F$, with $h(.)$ equalling the identity link. The orthocomplement of the nuisance tangent space for $\eta_1, \eta_2, \eta_3$ in this model consists of all mean zero functions of $(Y, X, V, S, C)$ (with $C$ the cluster index), whose expected value, given $(Y, X, V, S, b)$, is an element of $\left(\Lambda_1^F + \Lambda_2^F + \Lambda_3^F\right)^{\perp}$. Thus, for $\epsilon(\omega) = Y - \beta X - \gamma V$, $(\Lambda_1 + \Lambda_2 + \Lambda_3)^{\perp}$ equals $\{[D - E(D\,|C)]\,\epsilon(\omega) : D = d(X, V, S) \text{ arbitrary}\}$ because $b \amalg (X, V, S)|C$ so that the within-cluster sample average $E(D|C) = E(D|C, b)$ and $(X, V, S, Y) \amalg C|b$ (by the fact that clusters are identically distributed conditional on $b$) so that $E\{E(D|C, b)|X, V, S, Y, b\} = E(D|b)$. Note that we replaced $\epsilon(\omega, \nu, \eta_3)$ by $\epsilon(\omega) = Y - \beta X - \gamma V$ because $\{D - E(D\,|C)\}\{\alpha + \delta S + g(b; \eta_3)\}$ has mean zero regardless of $(\omega, \nu, \eta_3)$ and hence is orthogonal to the tangent space for $\omega$. This proves part 1b of Theorem 3.

Consider now $\mathcal{A}^F$ with $h(x) = \exp(x)$. Then (3.16) is satisfied iff $E[D\,\exp\{\alpha + \beta X + \gamma V + \delta S + g(b; \eta_3)\}|b] = 0$, so that $\left(\Lambda_1^F + \Lambda_2^F + \Lambda_3^F\right)^{\perp}$ equals

$$\{D\,\epsilon(\omega, \nu, \eta_3) : D = d(X, V, S, b) \text{ arbitrary satisfying}$$

$$E\{D\,\exp(\beta X + \gamma V + \delta S)\,|b\} = 0\}$$

It follows using similar arguments as in the previous paragraph that for the exponential link $(\Lambda_1 + \Lambda_2 + \Lambda_3)^{\perp}$ equals

$$\{[D - E(D|C)]\,Y\exp(-\beta X - \gamma V) : D = d(X, V, S; \omega) \text{ arbitrary}\}$$

This proves part 2b of Theorem 3.

Consider finally the full data model $\mathcal{A}^F$ with $h(.)$ equaling the inverse logit link. Then we must determine those functions $D = d(X, V, S, b; \omega)$ that satisfy

$$E\left(\frac{D\ \exp(\beta X + \gamma V + \delta S)}{[1 + \exp\{\alpha + \beta X + \gamma V + \delta S + g(b; \eta_3)\}]^2}|b\right) = 0$$

The orthocomplement of the nuisance tangent space for $\eta_1, \eta_2, \eta_3$ in the observed data model $\mathcal{A}$ with $h(.)$ equaling the inverse logit link, contains only 0 because there is no function $d(X, V, S; \omega)$ of $(X, V, S)$ that differs from zero and satisfies the above equality for any $b$. It follows that no root-$n$ estimators of $\beta$ and $\gamma$ exist in the observed data model $\mathcal{A}$ when $h(.)$ is the inverse logit link. This proves part 3 of Theorem 3.

The proof of parts 1a and 2a of Theorem 3 follows the lines of Appendix B of Robins, Rotnitzky and Zhao (1994) with $b_i(\gamma; d, \phi)$ replaced by $b_i(d, \omega) \equiv \mathbf{G}_i^{c1}(\mathbf{d}, \boldsymbol{\omega})$ when $h(.)$ equals the identity link and $b_i(d, \omega) \equiv \mathbf{G}_i^{c2}(\mathbf{d}, \boldsymbol{\omega})$ when $h(.)$ equals the inverse log link.

# Appendix 3.A3: proof of Theorem 4

To calculate the efficient score for $\omega$, we project the observed data score $S_\omega$ for $\omega$ corresponding to the true parametric submodel (Bickel et al., 1993) onto $(\Lambda_1 + \Lambda_2 + \Lambda_3)^\perp$. Let us define $\tilde{Y}(\omega) = Y - L\omega$ when $h(.)$ is the identity link and $\tilde{Y}(\omega) = Y \exp(-L\omega)$ when $h(.)$ is the inverse log link. It can easily be deduced from

$$E\{\tilde{Y}(\omega)|L,S,C\} = E\{\tilde{Y}(\omega)|S,C\}$$

where $C$ is the cluster index, that

$$E\left\{S_\omega\left(\tilde{Y}(\omega) - E\{\tilde{Y}(\omega)|S,C\}\right)|L,S,C\right\} = \{L - E(L|S,C)\}E\{\tilde{Y}'(\omega)|S,C\}$$

where we define $\tilde{Y}'(\omega) = 1$ when $h(.)$ is the identity link and $\tilde{Y}'(\omega) = \tilde{Y}(\omega)$ when $h(.)$ is the inverse log link. The efficient score for $\omega$ is a function

$$(d_0 - \bar{d}_0)\{\tilde{Y}(\omega) - E\{\tilde{Y}(\omega)|S,C\}\}$$

where $d_0 \equiv d_0(L,S,C)$ is such that for arbitrary $d \equiv d(L,S,C)$

$$\begin{aligned} 0 &= E\left[\left\{S_\omega - (d_0 - \bar{d}_0)\{\tilde{Y}(\omega) - E\{\tilde{Y}(\omega)|S,C\}\}\right\}\right. \\ &\qquad \left.(d - \bar{d})\{\tilde{Y}(\omega) - E\{\tilde{Y}(\omega)|S,C\}\}\right] \\ &= E\left[\left\{\{L - E(L|S,C)\}E\{\tilde{Y}'(\omega)|S,C\}\right.\right. \\ &\qquad \left.\left.-(d_0 - \bar{d}_0)Var\{\tilde{Y}(\omega)|L,S,C\}\right\}(d - \bar{d})\right] \end{aligned}$$

Assuming that $Var\{\tilde{Y}(\omega)|L,S,C\}$ is a constant variance matrix, it follows that for arbitrary $d \equiv d(L,S,C)$

$$0 \quad = \quad E\left[\left\{\{L - E(L|S,C)\}E\{\tilde{Y}'(\omega)|S,C\} - (d_0 - \bar{d}_0)Var\{\tilde{Y}(\omega)|L,S,C\}\right\}d\right]$$

from which we find

$$d_0 = LE\left\{\tilde{Y}'(\omega)|S,C\right\}Var^{-1}\{\tilde{Y}(\omega)|L,S,C\}$$

This proves Theorem 4.

# Appendix 3.A4: comparison CGEE- versus NK-approach

Using Taylor series expansion, it can be shown that the (efficient) estimating equations for $\beta_w$ in the observed data model (3.1) with $h(.)$ the identity link, accounting for estimation of the nuisance parameters $\beta_0$ and $\beta_b$, equal

$$\sum_i \left[ d_{wi} - E\left\{ d_{wi}\left(1, \overline{X}_i\right)\right\} E^{-1}\left\{ d_{bi}\left(1, \overline{X}_i\right)\right\} d_{bi}\right] \epsilon_i,$$

with $d_{wi} = (X_i' - \overline{X}_i)Var^{-1}(Y_i|X_i)$, $d_{bi} = \left(1, \overline{X}_i\right)' Var^{-1}(Y_i|X_i)$ and $\epsilon_i = Y_i - \beta_0 - \beta_w(X_i - \overline{X}_i) - \beta_b \overline{X}_i$. Suppose for the purpose of illustration that the true data-generating model equals

$$E(Y_{ij}|X_i, b_i) = \beta X_{ij} + b_i \tag{3.17}$$

where $b_i$ is a random effect which may be correlated with $X_i$. Then it follows from the proof of Part 1 of Theorem 3 that consistent estimators for the within-cluster effects under model (3.17) can be obtained from estimating equations of the form $\sum_i d_i \epsilon_i$ with $E(d_i|C_i) = 0$ and $\epsilon_i = Y_i - \beta X_i$, where $C_i$ is the index of cluster $i$. The condition $E(d_i|C_i) = 0$ is satisfied for the above equations since it can easily be shown that $E(d_{wi}|b_i) = 0$ and $E\left\{ d_{wi}\left(1, \overline{X}_i\right)\right\} = 0$ when the variance matrix $Var(Y_i|X_i)$ does not depend on $X_i$ other than through cluster-level summaries of $X_i$. We conclude that the NK-approach yields estimators within our class of CGEE-estimators and hence yields consistent estimators of within-cluster effects in the presence of unmeasured cluster-level confounders under the identity link and

homoscedasticity (w.r.t. $X$). Efficient estimators under model (3.1) with $h(.)$ the identity link are also efficient under model (3.17). The key to showing this is that they are CGEE estimators and hence consistent regardless of whether one consistently estimates the cluster-specific part of the model (e.g. $\beta_b$). It then follows from Newey and McFadden (1994) that their asymptotic distribution remains unchanged if one fixes $\beta_b = \beta_w$ (i.e. if one conducts inference under model (3.17)).

Efficient estimators for $\beta$ in model (3.17) can also be obtained by fitting model

$$E(Y_{ij} - \bar{Y}_i | X_i) = \beta(X_{ij} - \bar{X}_i) \tag{3.18}$$

This can be seen from the following arguments. It is immediate that model (3.17) implies the above model (3.18). That also the reverse is true can be seen upon writing $E(Y_{ij}|X_i) = \mu(X_i)$; under model (3.18), $\mu(X_i)$ must satisfy $\mu(X_i) - \bar{\mu}(X_i) = \beta(X_{ij} - \bar{X}_i)$. Writing $\mu(\bar{X}_i) = \beta\bar{X}_i + b_i$ without loss of generality, where $b_i$ is now a function of $X_i$, we find that $\mu(X_i) = \beta X_{ij} + b_i$. It follows that models (3.17) and (3.18) impose the same restrictions on the observed data law. Furthermore, the parameter $\beta$ in these models is the same functional

$$\beta = \frac{\mathrm{Cov}(Y_{ij} - \bar{Y}_i, X_{ij} - \bar{X}_i)}{\mathrm{Var}(X_{ij} - \bar{X}_i)}$$

of the observed data law under both models. Consequently, the set of CAN estimators for $\beta$ under model (3.17) equals the set of CAN estimators for $\beta$ under model (3.18). In particular, efficient estimators for $\beta$ under both models are asymptotically equivalent. By similar arguments, efficient estimators for $\beta$ in model (3.17) can also be obtained by fitting a model for the changes $Y_{ij} - Y_{i1}$.

Finally, we show that the NK-approach yields no consistent estimators of within-cluster effects $\beta$ under model (3.2) when $h(.)$ is the inverse log link. Key to this proof is the fact that the estimating functions under this approach do not belong to the class of CGEE estimating functions. Suppose that $E(Y_{ij}|X_i, b_i) = \exp(\beta X_{ij} + b_i)$ where $b_i$ is possibly correlated with $X_i$. Then fitting model $E(Y_{ij}|X_i) = \exp\{\beta^*(X_{ij} - \bar{X}_i)\}$ yields possibly

inconsistent estimates for $\beta$ in the above model because the estimating functions may have mean

$$E\left(d_j(X_i)\left[Y_{ij} - \exp\{\beta(X_{ij} - \bar{X}_i)\}\right]\right)$$
$$= E\left(d_j(X_i)\exp\{\beta(X_{ij} - \bar{X}_i)\}\left[\exp(\alpha + \beta\bar{X}_i + b_i) - 1\right]\right)$$

different from zero. Indeed, let $d_j(X_i) = X_{ij} - \bar{X}_i$ and suppose that $X_{ij} - \bar{X}_i | \bar{X}_i \sim N(0, \sigma_i^2)$ within the $i$th cluster. Then, using the moment generating function of normally distributed variates, we find that the cluster-average of $d_j(X_i)\exp\{\beta(X_{ij} - \bar{X}_i)\}$ equals $\exp(\sigma_i^2\beta^2/2)\sigma_i^2\beta$ which differs from zero for $\beta \neq 0$, and likewise the mean of $\exp(\alpha + \beta\bar{X}_i + b_i) - 1$ is not guaranteed to be zero. Allowing for an intercept in the model and further adjusting for the cluster mean provides no solution because we are precisely looking for estimators which are unaffected by adjustment for cluster-level variables. We conclude that the class of estimating functions under model (3.1) with $h(.)$ the inverse log link contains functions which are biased under the data-generating model (3.2) and, hence, may yield inconsistent estimators of $\beta$ under model (3.2). The proof further shows that bias under the NK-approach gets more severe with increasing effect sizes $\beta$ and within-cluster exposure variation $\sigma_i^2$. Hypothesis tests for the absence of an effect will preserve their nominal $\alpha$-level (as there is no bias in the absence of an effect, i.e. $\beta = 0$), but may be less powerful than tests obtained using CGEEs.

Intuitively, the reason why the NK-approach fails under nonlinear link functions can be seen as follows. Consider first model (3.17) with identity link. This model implies model (3.18), from which all cluster-level confounders have been removed. It follows that valid estimates for the effect $\beta$ can be obtained by fitting the latter model, even in the presence of cluster-level confounding. The NK-approach is a slight variation of this model whereby $E(\bar{Y}_i | X_i, b_i)$ is replaced by the fitted value from a regression model

$$E(\bar{Y}_i | X_i, b_i) = \beta_b\bar{X}_i + b_i$$

In fact, the NK-approach under the identity link is an immediate application of G-estimation whereby the exposure in a linear model is replaced

by its residual from a regression on confounders (i.e. on the cluster index). Suppose now that $E(Y_{ij}|X_i, b_i) = \exp(\beta X_{ij} + b_i)$. This model cannot be rewritten in terms of within-cluster exposures $X_{ij} - \bar{X}_i$ such that the cluster-level confounders $b_i$ disappear. For instance, while the cluster-level confounder $b_i$ can be removed through the transformation

$$\frac{E(Y_{ij}|X_i, b_i)}{E(\bar{Y}_i|X_i, b_i)} = \frac{\exp(\beta X_{ij})}{\sum_{k=1}^{n_i} \exp(\beta X_{ik})/n_i}$$

this does not imply a loglinear model involving within-cluster exposures $X_{ij} - \bar{X}_i$ due to the nonlinearity of the link function.

# Chapter 4

# Introduction to direct effect estimation

**Summary**

In this chapter, we will outline the difficulties in inferring direct exposure effects. We start by introducing concrete definitions of direct effects in terms of potential outcomes notation. We compare these different direct effects definitions and we discuss their usefulness and relevance in different settings. Then, the problem with inferring direct effects is explained by using a causal DAG. Besides explaining the difficulties in inferring these effects, we will show with a case study that in certain settings, direct effect estimates can be easily obtained. In addition, we will investigate whether structural equation models (SEM) (introduced in Chapter 2) can be used to estimate direct effects and finally, we give an brief overview of recent literature on direct effect estimation. In the next chapters, we will develop solutions to address the difficulties described in this chapter.

## 4.1 Motivation for direct effects

The causal effect of an exposure on an outcome may manifest itself through various causal paths. For example, advanced maternal age at pregnancy may increase the risk of hypertension or diabetes during pregnancy,

which adversely affects birth weight. Advanced maternal age may also increase the risk of twins, which again adversely affects birth weight. Besides these indirect effects, maternal age may directly affect birth weight. In fact, in most cases the total effect of an exposure on an outcome (i.e. through all causal paths) is virtually always a combination of direct and indirect effects. Researchers' interest is frequently not only in this total effect, but also in the effect that is not mediated through intermediate variables, i.e. the direct effect of the exposure on the outcome. The following examples motivate this.

(*Example 1* Verstraelen, Goetgeluk et al., 2005) Artificial reproductive technologies (ART) can have an effect on perinatal outcomes for twins through various causal pathways. For example, the occurrence of dizygotic (DZ) twins is much larger in the group of babies conceived through ART than in the group conceived naturally. Because DZ twins have better perinatal outcomes than monozygotic (MZ) twins, this may induce an indirect effect of ART on perinatal outcomes through twin zygosity. In Section 4.4, besides inferring a total causal effect, we infer the effect of ART on perinatal outcomes that is not mediated through type of twinning (MZ or DZ).

(*Example 2* Vansteelandt, Goetgeluk et al., 2008) Lyon et al. (2004) find SNPs in the IL10 gene to be associated with both body mass index (BMI) and forced expiratory volume (FEV). Because BMI and lung function are themselves associated (Oliveti et al., 2006), this raises the question whether a genetic effect found on one of these phenotypes is actually an indirect effect through the other (intermediate) phenotype. In Chapter 5, we infer the genetic effect of the SNPs of interest on FEV which is not mediated through BMI.

(*Example 3* Goetgeluk et al., 2008) De Sutter et al. (2006) estimate the effect of single versus double embryo transfer (SET versus DET) on birth weight. They observe birth weights to be 120 grams (95% confidence interval [44;197]) lower on average in singletons born after double than single embryo transfer. In response to criticism that the analysis was not adjusted for gestational age, Delbaere et al. (2007b) argue that such adjustment would remove a possible indirect effect of SET/DET on birth weight through gestational age, and that this could even introduce selection bias

(see Section 4.3). At the same time, the debate raises the question whether the effect of SET/DET on birth weight is entirely mediated through gestational age. In Chapter 6, we will discuss this in more detail and infer the effect of SET/DET on birth weight that is not mediated through gestational age.

(*Example 4* Rosenblum et al., 2007) When investigating the effect of diaphragm and lubricant gel use on HIV infection on the basis of the randomized Methods for Improving Reproductive Health in Africa (MIRA) trial, Rosenblum et al. (2007) observe much lower reported condom use in the treatment arm than in the untreated arm. This makes it difficult to answer important public health questions solely on the basis of the intention-to-treat analysis. In view of this, they estimate the effect on HIV infection of assignment to diaphragm and lubricant gel use, were all participants to consistently use condoms during all sex acts. That is, they infer the effect of treatment on HIV infection which was not mediated through condom use.

We will now introduce concrete definitions of direct effects in terms of potential outcomes notation.

## 4.2 Definitions of direct effect

We define $Y_x$ as the potential outcome a given subject would have had if he/she received exposure $X = x$. The average total effect of exposure $x$ versus 0 is then an average contrast between $Y_x$ and $Y_0$, e.g. $\mathrm{E}(Y_x - Y_0)$. In a causal diagram, we may consider a number of intermediate variables/mediators $K$, which are affected by $X$ and in turn have an effect on $Y$. A direct effect of $X$ on $Y$ other than through modifying an intermediate variable $K$, expresses the effect of, say, exposure $x$ versus 0 had the intermediate variable $K$ remained unchanged. This requires us to introduce potential outcomes following joint exposures $x$ and $k$. Specifically, define $Y_{xk}$ as the potential outcome which a given subject would have experienced under exposure $X = x$ and a fixed value $k$ for the intermediate variable $K$. Throughout, we make the consistency assumption that this potential outcome equals the observed outcome for subjects for whom $x$

and $k$ correspond to the observed values for the exposure and the intermediate variable, respectively. Using this notation, a number of definitions for direct effects have been proposed in the literature.

1. **Controlled direct effect** (Robins, 1999b; Pearl, 2001; Robins and Greenland, 1992, Petersen, Sinisi and van der Laan, 2006; Didelez, Dawid and Geneletti, 2006; Goetgeluk et al., 2008)
   The individual controlled direct effect of setting exposure $X$ to $x$ (versus setting $X$ to 0) on outcome $Y$, when holding $K$ fixed at a value $k$, is defined as the contrast $Y_{xk} - Y_{0k}$ between the two potential outcomes $Y_{xk}$ and $Y_{0k}$ for the same subject. The (population) controlled direct effect is defined as the average of the individual controlled direct effects taken over all subjects, i.e.

   $$E(Y_{xk} - Y_{0k})$$

   It expresses how much the expected outcome would change if the exposure $X$ changed from $x$ to 0, but the intermediate variable $K$ were kept uniformly fixed at a given $k$.

2. **Natural direct effect** (Pearl, 2001; Robins and Greenland, 1992; Petersen, Sinisi and van der Laan, 2006; Didelez, Dawid and Geneletti, 2006)
   The individual natural direct effect on outcome $Y$ of setting exposure $X = x$ (versus $X = 0$), when holding $K$ fixed at the value $K_0$, which the intermediate would have had were the subject not exposed, is defined as the contrast $Y_{xK_0} - Y_{0K_0}$ between the two potential outcomes $Y_{xK_0}$ and $Y_{0K_0}$ for the same subject. The main difference with controlled direct effects is that the value $K_0$ can be different for all subjects (that is, $K_0$ is a random variable). The (population) natural direct effect is defined as the average of the individual natural direct effects taken over all subjects, i.e.

   $$E(Y_{xK_0} - Y_{0K_0})$$

   It expresses how much the expected outcome would change if the exposure $X$ changed from 0 to $x$, but its effect on the intermediate variable $K$ were blocked.

3. **Principal stratification direct effect** (Rubin, 2004; Frangakis and Rubin, 2002)
   Principal stratification direct effects measure the average causal effect of setting exposure $X = x$ (versus $X = 0$) among subjects for whom the intermediate variable was not affected by $X$; that is,

   $$E(Y_x - Y_0 | K_x = K_0)$$

   where $K_x$ is the potential (counterfactual) value of the intermediate variable $K$ corresponding to setting $X = x$.

4. **Standardized direct effects** (Didelez, Dawid and Geneletti, 2006)
   Standardized direct effects are obtained by averaging controlled direct effects over a chosen density function $f^*(K)$ for the mediator, which does not depend on the exposure; that is,

   $$\int E(Y_{xk} - Y_{0k} | K = k) f^*(K = k) dk$$

   for a chosen density function $f^*(K)$. Equivalently, they can be written as the expected contrast

   $$E(Y_{xK^*} - Y_{0K^*})$$

   where $K^*$ is a random draw from the distribution $f^*(K)$, independently of $x$. Natural direct effects form a special case obtained by setting $f^*(K) = f(K_0)$.

We illustrate the difference between these definitions using Example 3. The direct effect of SET ($X = 0$) versus DET ($X = x$) on birth weight, not mediated through gestational age translates as follows under the different direct effect definitions:

1. The controlled direct effect of SET/DET on birth weight expresses how different the average birth weight would have been had the study population uniformly experienced SET versus had they uniformly experienced DET, but gestational age were fixed for all women at the same value, e.g. 266 days (i.e. 38 weeks). By fixing gestational age to

be the same for all women, the indirect effect of SET/DET through gestational age is blocked.

For this direct effect parameter to be well defined, it must be possible, at least in principle, to imagine interventions that would fix gestational age to be the same for all women and that do not affect birth weight in any other way (Frangakis and Rubin, 2002). While such interventions might technically be approximately possible, this direct effect parameter may not be the most relevant one from a public health perspective because one would never consider doing an intervention such that all women have the same gestational age.

2. The natural direct effect of SET/DET on birth weight expresses how much the average birth weight would change if the study population would uniformly experience DET instead of SET, but gestational age remained unchanged. Under this definition, we also block the indirect effect of SET/DET through gestational age, but allow for different women to have different gestational ages so as to obtain a more natural definition of direct effects. Note however that it is not technically possible to imagine interventions that would realize this. This is because, even if one could fix gestational age at a given value, the problem is that for those women undergoing DET, it is unclear what their gestational age would have been, had they experienced SET.

3. The principal stratification direct effect of SET/DET on birth weight is the difference in average birth weight between SET and DET for women whose gestational age is the same under SET as under DET. Because this direct effect parameter only relates to a small subgroup of all subjects, inferences for principal stratification direct effects must deal with a sparse data problem (Robins, Rotnitzky and Vansteelandt, 2007). Furthermore, while this effect estimand is well defined, even when one cannot imagine interventions that fix gestational age, it may have limited public health relevance (a) because one can never identify women whose gestational age would be the same under SET as under DET; and (b) because this group of women to which the direct effect estimand relates may be a very small subset of the pop-

ulation.

4. The standardized direct effect of SET/DET on birth weight is the difference in average birth weight had the study population uniformly experienced SET versus DET, but gestational age for each subject were fixed at a certain value which may be different for every subject. It could for example be the average value for women of the same age who had spontaneous conception.

Each of the different direct effect definitions have their advantages and disadvantages. They may be useful in different situations/contexts. The controlled direct effect is meaningful when one can imagine realistic interventions that fix the value of the intermediate variable to be the same for all subjects. When the restriction for all subjects to have the same value for the intermediate is too restrictive, natural and standardized direct effects become more meaningful. For example, Petersen et al. (2006) infer the natural direct effect of air pollution on lung function, not mediated through use of respiratory (rescue) medication. They do so because it is not meaningful to imagine fixing the level of this medication to be the same (e.g. no use of the medication) for the entire study population. This is because there are likely children in the study population whose underlying respiratory illness is such that they always need this medication, regardless of air pollution level. They henceforth fix the level of medication use for each child at the level it would have been if there were no air pollution. A drawback of inference for natural direct effects is that stronger assumptions are required because the potential use of medication in the absence of air pollution is not directly observable.

In certain situations, natural, controlled and standardized direct effects are equivalent. This is so when the exposure and the intermediate variable do not interact to affect the outcome (Robins and Greenland, 1992). That is, when the controlled direct effect satisfies the no-interaction assumption that

$$E(Y_{xk} - Y_{0k}) = E(Y_{xk'} - Y_{0k'})$$

for all $k' \neq k$ in the support of $K$. This and the fact that natural direct effects are more difficult to infer, explains why inference for controlled direct

effects is of interest, despite natural direct effects often being more relevant. Moreover, controlled direct effects can easily be estimated using standard regression techniques, under specific assumptions (see next section). To estimate natural direct effects, simple regression cannot be used and extra assumptions are needed (Petersen et al., 2006). In particular, note that natural direct effects average controlled direct effects

$$E(Y_{xK_0} - Y_{0K_0}) = \int E(Y_{xk} - Y_{0k}) f(K_0 = k) dk$$

provided that the untestable assumption holds that (Petersen et al., 2006)

$$E(Y_{xk} - Y_{0k} | K_0 = k) = E(Y_{xk} - Y_{0k})$$

Note from the above result that identifying natural direct effects additionally requires causal assumptions to identify $f(K_0)$.

The advantage of principal stratification direct effects is that they do not consider interventions on the intermediate variable and are thus always well defined in the context of randomized studies (in which there are no unmeasured common causes of exposure and outcome). However, for continuous intermediate variables, it is likely that the principal stratum of subjects whose level of the intermediate variable was not affected by the exposure, will be a set of probability mass zero (Robins, Rotnitzky and Vansteelandt, 2007). This implies that inferences may become unstable and that this parameter is meaningful only for a very small subset of the population. Therefore, this direct effect parameter is mostly used when the intermediate is qualitative.

In the remainder of this thesis, we focus on estimating controlled direct effects because they are more easily obtained, require less assumptions and because common language about direct effects usually reflects controlled direct effects. From now on, we will use 'direct effects' and 'controlled direct effects' interchangeably.

## 4.3 The problem with inferring direct effects

Suppose we are interested in estimating the direct effect of exposure $X$ on outcome $Y$ that is not mediated through the intermediate variable $K$.

A causal DAG that illustrates this scenario is shown in Figure 4.1. The causal DAG represents a setting of a randomized treatment $X$ which may have both a direct and indirect effect on outcome $Y$. Even in this ideal experimental setting, there may be unmeasured factors $U$ jointly affecting the intermediate variable and the outcome. We will see that the existence of such variables complicates direct effects inference, even in randomized trials (Cole and Hernan, 2002; Rosenbaum, 1984).

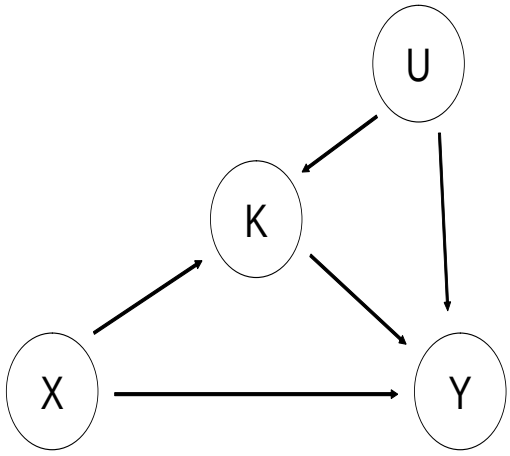


Figure 4.1: *Causal diagram representing the (direct) effect of exposure $X$ on outcome $Y$ which is not mediated through intermediate variable(s) $K$*

To examine whether the direct effect of $X$ on $Y$ is identifiable, we use d-separation. In the absence of a direct effect (i.e., when we ignore the arrow between $X$ and $Y$) the DAG in Figure 4.1 reveals an open path from $X$ to $Y$ through $K$. This path can be blocked by adjusting the association between $X$ and $Y$ for $K$. Without such adjustment, the path from $X$ to $Y$ through $K$ and $U$ was blocked because $K$ is a collider along that path. However, this path becomes unblocked upon adjusting for $K$. In order to close this path again, we need to additionally adjust for $U$. When all variables $U$ are measured, this causes no problems and the direct effect

can be estimated. Often however, not all common causes of $K$ and $Y$ are measured or known, and thus, the estimated direct effect, as obtained via traditional regression adjustment for $K$, is biased in that case.

For the remainder of this chapter, we examine under what conditions the direct effect of $X$ on $Y$ can be identified. In the following chapters, we will develop novel direct effect estimators which require fewer assumptions than direct effect estimators obtained via traditional regression methods.

## 4.4 Case study: inferring the direct effect of ART on perinatal health other than through zygosity

The previous development shows that, under certain conditions, direct effects can be relatively easily inferred through traditional regression adjustment for the intermediate variable. This is the case when the intermediate variable $K$ does not share any other causes with the outcome than the exposure. Traditional regression adjustment is also valid when the common causes $L$ of the intermediate variable and the outcome are measured and adjusted for, and in addition they either share no unmeasured common causes $U$ with the outcome or are not affected by the exposure. This will be explained in more detail at the end of this section. First, we will illustrate this through the following example.

Increasingly more couples attempting pregnancy fail to conceive naturally within a year and seek help through subfertility treatments (Taylor, 2003). Efforts to increase the success rates of subfertility treatment have been accompanied by an insidious rise in the rate of multifetal pregnancies (Blondel et al., 2002). In the face of this multiple birth epidemic, and despite widespread concern about the effects of medically aided conception on perinatal outcome, few studies have investigated outcomes in twins (Blondel et al., 2003), and largely conflicting results have been reported (Helmerhorst et al., 2004). Twins tend to fare considerably worse than singletons, with much higher rates of perinatal mortality, neonatal morbidity, and long term neurological impairment (Blondel et al., 2002). Adverse

pregnancy outcome in turn relates to the high prevalence of preterm birth among twins (Derom et al., 2005; Loos et al., 1998). Whether subfertility treatment also impinges on gestational length in twins, as has been established among singletons (Helmerhorst et al., 2004; Jackson et al., 2004), is unclear, as is the extent to which type of twinning interferes with perinatal outcome after subfertility treatment (Machin, 2004).

To investigate whether there is a direct effect of ART on perinatal outcomes that is not mediated through zygosity, we first draw a DAG containing the exposure (ART) and outcome (perinatal outcomes) of interest. Here, three types of ART are being considered: ovulation induction, in vitro fertilization and intracytoplasmic sperm injection. Since there is an interest in investigating to which extent type of twinning (i.e. zygosity) affects perinatal outcomes after subfertility treatment, we add zygosity to the DAG. Finally, we add all common causes of all pairs of variables to the DAG. Measured common causes of ART and perinatal outcomes are parity (i.e. the number of older children), maternal age and year of birth. In our analysis below, we assume that there are no other (possibly unobserved) common causes of ART and perinatal outcomes.

Direct effects inference is simplified in this setting because there are few processes acting upon zygosity. Besides type of conception (naturally or through ART) and maternal age[1], we believe there are few processes that may have an influence on type of twinning. Thus, we draw an extra arrow from maternal age to zygosity and assume that there are no remaining common causes of zygosity and perinatal outcomes. The resulting causal DAG is shown in Figure 4.2.

Note that the analysis is restricted to twins. This may introduce selection bias since ART makes it more likely to have twins[2] and thus restricting the analysis to twins involves adjustment for a post-treatment measurement. As explained in the previous section, such adjustment is problematic whenever there exist risk factors for perinatal outcomes which are also as-

[1]The older the mother at the time of conception, the more chance of having a DZ twin (Hoekstra et al., 2008).

[2]This is so for hormone use, and was until recently also the case for IVF and ICSI, since two or more embryos were often transferred to the womb.

sociated with multiplicity (i.e. having a singleton or multiple birth). Such risk factors may well exist. For instance, this would be the case if certain genes that predispose people to having multiples, also affect perinatal outcomes. In that case, this unmeasured genetic variable, may introduce selection bias. Another possible risk factor is fertility of the mother. Likewise, maternal age affects multiplicity and perinatal outcomes, but this variable is not problematic since it is already accounted for in the analysis.

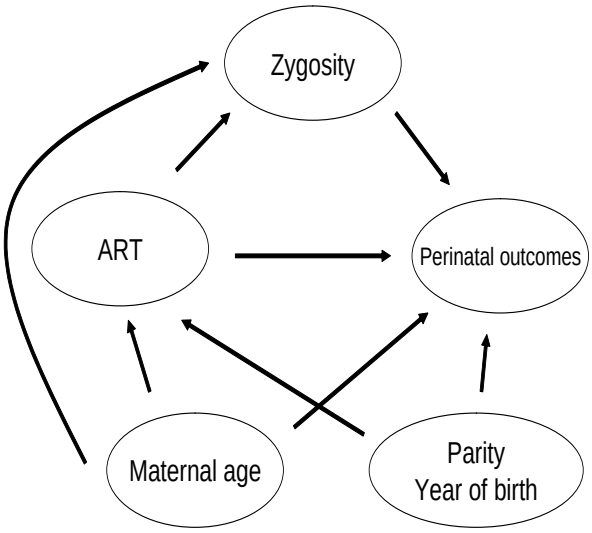


Figure 4.2: *Causal diagram for estimating direct effect of ART on perinatal outcomes*

We focus on two main perinatal outcomes, birth weight and gestational age. In a first analysis, the total effect (i.e. combination of direct effect and indirect effects) of ART on birth weight and gestational age is estimated. D-separation indicates that adjustment for the confounders parity, maternal age and year of birth is necessary. We estimate the total effect of ART on the risk of low birth weight and risk of preterm birth through marginal

logistic regression models, which are fitted using generalized estimating equations (Diggle et al., 1994) with exchangeable working correlation and different correlation allowed for monozygotic and dizygotic twins (in SAS version 9.1, with proc genmod). The results can be viewed in the first column of Table 4.1. In a second analysis, the goal is to estimate the direct effect of ART on the risk of low birth weight and on the risk of preterm birth, which is not mediated through zygosity. Since no other variables than ART and maternal age are assumed to affect zygosity, additionally adjusting for zygosity[3] in the analysis for inferring the direct effect of ART on perinatal outcomes opens the closed sequence of arrows that point to zygosity (zygosity is a collider), but does not induce bias because the path is closed by adjusting for maternal age. The results can be viewed in the second column of Table 4.1.

| | Preterm birth | |
|---|---|---|
| | total effect OR | direct effect OR |
| OI vs. SC | 1.26 [1.07,1.50] | 1.46 [1.22,1.75] |
| IVF/ICSI vs. SC | 1.38 [1.15,1.66] | 1.73 [1.34,2.00] |
| | Low birth weight | |
| | total effect OR | direct effect OR |
| OI vs. SC | 1.07 [0.93,1.24] | 1.29 [1.10,1.50] |
| IVF/ICSI vs. SC | 1.09 [0.93,1.24] | 1.32 [1.12,1.55] |

Table 4.1: *Odds ratios for preterm birth and low birth weight for total and direct effect of ART (ovulation induction (OI), IVF/ICSI or spontaneous conception (SC))*

We find that the odds of preterm birth is 1.26 (95%CI [1.07,1.50]) times higher for ovulation induction compared to spontaneous conception and 1.46 (95%CI [1.22,1.75]) times higher when fixing zygosity. The odds of preterm birth is 1.38 (95%CI [1.15,1.66]) times higher for IVF/ICSI versus spontaneous conception and 1.73 (95%CI [1.34,2.00]) times higher when fixing zygosity. All of these odds ratios are significantly different from 1 and thus, represent a systematic difference between all types of conception.

[3]next to maternal age, parity and year of birth

The total effect odds ratios (1.26, 95%CI [1.07,1.50], for ovulation induction versus spontaneous conception and 1.38, 95%CI [1.15,1.66], for IVF/ICSI versus spontaneous conception), however, are smaller than the direct effect odds ratios. This is due to the mediation of the direct negative effect of ART on gestational age by the indirect beneficial effect of ART on gestational age since there are more dizygotic twins after ART who, on average, have an older gestational age than monozygotic twins. The odds of low birth weight is only significantly higher (e.g. 1.29 (95%CI [1.10,1.50]) times and 1.32 (95%CI [1.12,1.55]) times respectively) for ovulation induction and IVF/ICSI compared to spontaneous conception for the direct effect, not for the total effect (1.07, 95%CI [0.93,1.24], for ovulation induction versus spontaneous conception and 1.09, 95%CI [0.93,1.24], for IVF/ICSI versus spontaneous conception). This is again due to the mediation of the direct negative effect of ART on birth weight by the indirect beneficial effect of ART on birth weight through zygosity.

We conclude that infertility is associated with a quite large negative effect on gestational age and this effect is somewhat larger for IVF/ICSI than for ovulation induction (Verstraelen et al., 2005). Our results suggest that, in contrast to the general belief, twins born after ART have an additional risk of preterm birth over and above the risk implied by being a twin. However, this increased risk diminishes due to the large number of dizygotic twins after ART. Thus, when zygosity is not taken into account, the direct effect of ART on gestational age is underestimated. These results show the importance of correctly registering the type of zygosity and conception in studies of fertility treatment.

The above example is a simple illustration of direct effects estimation, where the simplicity results from few processes acting upon the intermediate variable. In most cases, however, inferring a direct effect is not as easy as in this case study. Problems arise when one (or more) of the measured confounders of the effect of the intermediate variable on the outcome, is itself affected by the exposure and, in addition, there exist prognostic factors of the outcome which are associated with this confounder. This situation is shown in the DAG in Figure 4.3. D-separation shows that additionally adjusting for such measured confounders $L$ of the intermediate and out-

come may then induce bias. It may do so by opening the path from $X$ to $Y$ through $L$ and $U$. Thus, when there are unmeasured confounders $U$ affecting $L$ and $Y$, simple regression adjustment does not give the causal effect of interest.

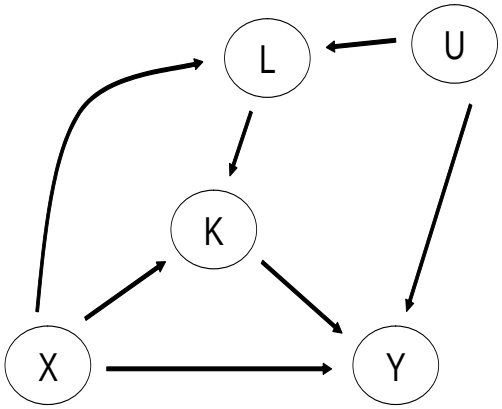


Figure 4.3: *Causal diagram for estimating direct effect of exposure $X$ on outcome $Y$ in presence of measured confounders $L$ and unmeasured confounders $U$*

The DAG in Figure 4.3 may seem complex at first sight, but is nonetheless representative of many realistic situations. This is so because the intermediate variable $K$ arises post treatment $X$ and thus it is likely that some of the confounders for the association between $K$ and outcome also arose only post treatment. In Example 2, for instance, height is a measured confounder for the association between body weight (the intermediate variable) and lung function (the outcome), and may itself be genetically affected. In Example 3, zygosity and multiplicity (e.g. twin/singleton) are measured confounders for association between the intermediate (gestational

age) and the outcome (birth weight), which may themselves be affected by SET/DET. In the example about the effect of ART on perinatal outcomes multiplicity is also a measured confounder for the association between the intermediate (zygosity) and perinatal outcomes. Note that in this example only twins are ascertained, and in Example 3 only singletons. Thus, these analyses implicitly adjust for multiplicity, which may induce bias, as explained above. In the former example, a possible solution is to constraint the analysis to the principal stratum of women who would have gotten a twin either way, after ART and after spontaneous conception. This way twin/singleton status is no longer adjusted for, but the counterfactual twin/singleton status is, which is no longer affected by type of conception. The same can be done in Example 3, by constraining the analysis to the principal stratum of women who would have gotten a singleton eitherway, after SET and after DET.

## 4.5 Structural equation models

### 4.5.1 Inferring direct causal effects via structural equation models

A frequently used alternative for direct effect inference is based on path diagrams and structural equation models. In this section, we evaluate how well these methods perform to estimate the direct effect of an exposure $X$ on an outcome $Y$ in the DAG in Figure 4.3 and under what assumptions they give valid estimates.

Remember from Chapter 2 that structural equation models parameterize a causal DAG using multiple linear models involving measured and possibly unmeasured variables. In particular, they postulate a multivariate normal distribution for the vector of measured and unmeasured variables, which satisfies the conditional independence assumptions imposed by the underlying DAG (see Chapter 2 for a more detailed explanation). For instance, the structural equation model corresponding to the diagram in Figure 4.3 is

$$\begin{aligned} X &= \alpha_0 + \epsilon_X \\ U &= \beta_0 + \epsilon_U \\ L &= \gamma_0 + \gamma_1 X + \gamma_2 U + \epsilon_L \\ K &= \delta_0 + \delta_1 X + \delta_2 L + \epsilon_K \\ Y &= \eta_0 + \eta_1 X + \eta_2 K + \eta_3 U + \epsilon_Y \end{aligned} \tag{4.1}$$

where $\epsilon_X$, $\epsilon_U$, $\epsilon_L$, $\epsilon_K$ and $\epsilon_Y$ are mutually independent, mean zero normal variates with variances $\psi_X$, $\psi_U$, $\psi_L$, $\psi_K$ and $\psi_Y$. If this model is correctly specified, unbiased estimates may be obtained of the path coefficients of interest, e.g. the direct effect parameters, provided that they are identifiable from the observed data distribution. It follows from the Causal Markov Assumption (see Chapter 1) that, assuming that the path diagram corresponding to the SEM is a causal diagram, SEM adjust for the correct set of variables in order to obtain the direct effect of $X$ on $Y$, and thus that the path coefficients represent direct causal effects (see also Pearl, 2000, for a more detailed argument).

### 4.5.2 Simulation study

In the previous sections, we have seen that it is important to consider the existence of unmeasured common causes of confounders $L$ and outcome $Y$, as these may bias traditional regression estimates whenever the confounder $L$ is affected by the exposure $X$. Structural equations must explicitly model such common causes. This may be problematic when these common causes are unmeasured as it may yield identifiability problems and because model misspecification then becomes difficult to detect. In this section, we investigate this through limited simulation studies under the DAG of Figure 4.3. All analyses were conducted in R (version 2.3.1), using function 'sem' to fit structural equation models and 'glm' to fit linear regression models.

*Simulation experiment 1*

We simulate 1000 data sets of size $n$ ($n = 100$ or $n = 1000$) corresponding to the path diagram in Figure 4.3, under the following models (corresponding to equations (4.1)):

$$\begin{aligned} X &= 2 + \epsilon_X \\ U &= \epsilon_U \\ L &= 3 + 2X + 2U + \epsilon_L \\ K &= 3 + 2X + 1.5L + \epsilon_K \\ Y &= 5 + 1.5X + 0.8K + 2U + \epsilon_Y \end{aligned} \tag{4.2}$$

where $\epsilon_X$, $\epsilon_U$, $\epsilon_L$, $\epsilon_K$ and $\epsilon_Y$ are mutually independent, mean zero normal variates with standard deviations 1.5, 1, 5, 0.5 and 1, respectively. We then repeat the simulation study with the effect of $X$ on $L$ set to zero. The previous theoretical results indicate that simple regression adjustment for $K$ and $L$ will be biased, except in the second simulation setting, where $L$ is not affected by $X$.

*Simulation experiment 2*

We simulate 1000 data sets of size $n$ ($n = 100$ or $n = 1000$) corresponding to the path diagram in Figure 4.3, under the following models (corresponding

to equations (4.1)):

$$\begin{aligned} X &= \epsilon_X \\ U &= \epsilon_U \\ L &= 3 + 2X + 2U + \epsilon_L \\ K &= 3 + X + 1.5L + \epsilon_K \\ Y &= 5 + 3X + 1.8K + 2U + \epsilon_Y \end{aligned} \tag{4.3}$$

where $\epsilon_X$, $\epsilon_U$, $\epsilon_L$, $\epsilon_K$ and $\epsilon_Y$ are mutually independent, mean zero normal variates with standard deviations 1.5, 1, 5, 0.5 and 1, respectively. Compared to the previous simulation setting, we have strengthened the direct effect of $X$ on $Y$ and weakened the effect of $X$ on $K$. This way, the indirect effect of $X$ on $Y$ is weakened and we expect that adjusting for $K$ will cause less bias than in the previous simulation. As before, we repeat the simulation study with the effect of $X$ on $L$ set to zero.

*Simulation experiment 3*

We simulate 1000 data sets of size $n$ ($n = 100$ or $n = 1000$) corresponding to the path diagram in Figure 4.3, under the following models (corresponding to equations (4.1)):

$$\begin{aligned} X &= 1 + \epsilon_X \\ U &= \epsilon_U \\ L &= 1 + 2X + U + \epsilon_L \\ K &= -0.5X + 0.5L + \epsilon_K \\ Y &= 5 + 2X + 0.5K + U + \epsilon_Y \end{aligned} \tag{4.4}$$

where $\epsilon_X$, $\epsilon_U$, $\epsilon_L$, $\epsilon_K$ and $\epsilon_Y$ are mutually independent, mean zero normal variates with standard deviations 0.5, 1, 3, 0.3 and 1, respectively. Compared to the previous simulation settings, the confounding effects of $U$ on $L$ and $Y$ are weakened now. As before, we repeat the simulation study with the effect of $X$ on $L$ set to zero.

*Simulation experiment 4*

In this simulation study we will examine the impact of misspecifying one of the linear models in the structural equation model. Specifically, we generate

a nonlinear relationship between $U$ and $Y$. As before, we simulate 1000 data sets of size $n$ ($n = 100$ or $n = 1000$) corresponding to the path diagram in Figure 4.3, under the following models (corresponding to equations (4.1)):

$$\begin{aligned} X &= \epsilon_X \\ U &= \epsilon_U \\ L &= 3 + 2X + 2U + \epsilon_L \\ K &= 3 + X + 1.5L + \epsilon_K \\ Y &= 5 + 3X + 1.8K + 2U + 1.5U^2 + \epsilon_Y \end{aligned} \tag{4.5}$$

where $\epsilon_X$, $\epsilon_U$, $\epsilon_L$, $\epsilon_K$ and $\epsilon_Y$ are mutually independent, mean zero normal variates with standard deviations 1.5, 1, 5, 0.5 and 1, respectively. Compared to Simulation 2, a quadratic effect of $U$ on $Y$ is added to the model for $Y$.

Again, we perform this simulation twice, once using the models above and once with the effect of $X$ on $L$ set to zero.

*Analysis*

In all analyses, $U$ is assumed to be unmeasured. The following analyses were considered.

SEM) We fit the structural equation model given in (4.1), with $\eta_1$ the direct effect parameter of interest. Because of identifiability problems, we follow standard practice by fixing certain path coefficients to 1. In particular, we set the effects of $U$ and its variance equal to 1.

LM) We ignore the unmeasured confounder $U$ by fitting the linear model

$$E(Y|X, K, L) = \kappa_0 + \kappa_1 X + \kappa_2 K + \kappa_3 L$$

to estimate the effect $\kappa_1$. D-separation indicates that this effect corresponds to the direct effect $\eta_1$ of interest only when $L$ is not affected by $X$.

UW) Finally, we apply a novel methodology that we develop in Chapters 5 and 6 to estimate direct effects (see Chapter 5 and the unweighted estimator in Section 6.2.4). This method overcomes the need to model the unmeasured confounders $U$ for $L$ and $Y$.

*Results*

The simulation results of Simulation experiment 1, 2 and 3 are shown in Table 4.2 for the case where there is an effect of $X$ on $L$ and in Table 4.3 for the case where there is no effect of $X$ on $L$. Table 4.4 gives the results for Simulation experiment 4. In Simulation experiment 3, in several iterations, the structural equation models failed to converge due to singularities. This was the case in 54 and 22 iterations in the setting with an effect of $X$ on $L$ for sample size 100 and 1000 respectively. For these iterations, the corresponding estimates for the unweighted estimator and the traditional regression adjustment estimator were excluded to maintain comparability with the SEM results. In case of no effect of $X$ on $L$, the structural equation models failed to achieve convergence in almost every iteration, which is why these results are not added to Table 4.3.

As expected the simple regression model only gives unbiased results when there is no effect of $X$ on $L$. The efficiency of the novel unweighted estimator is then comparable to that of traditional regression adjustment, suggesting that the protection guarantees offered by our estimators do not come at expense of imprecision in cases where traditional regression adjustment works. This is especially so at larger sample sizes (i.e. $n = 1000$). Moreover, Simulation experiment 3 illustrates that, in the setting with an effect of $X$ on $L$, the unweighted estimator may also seriously outperform the regression estimator in terms of MSE.[4]

In all simulations, the structural equations analysis gives biased results, even when all model assumptions are correct, as in Simulation experiment 1, 2 and 3. This is likely the result of fixing unknown parameter values at 1 to obtain an identifiable model. Note however that the standard errors of the structural equations estimates are much smaller than those obtained with the unweighted estimator or with the linear model. Because bias is not easily detectable, we believe however that the concern for bias trumps efficiency concerns. In particular, note that the confidence intervals obtained via structural equations analysis have very low coverage, in contrast

[4]Note that for the unweighted estimator, the estimated standard error deviates from the empirical standard error in several cases, thus, the assymptotics cannot always be trusted and bootstrap standard errors would be more appropriate.

to those obtained with the novel estimator. This is of concern, noting that confidence intervals are frequently used for decision making.

| Simulation experiment 1 | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | $n$ | Bias | 95% CI Bias | Estimated SE | Empirical SE | Coverage | MSE | MSE ratio |
| SEM | 100 | -0.35 | [-0.36;-0.34] | 0.19 | 0.20 | 0.55 | 0.40 | 5.63 |
| LM | 100 | -0.33 | [-0.38;-0.27] | 0.88 | 0.91 | 0.91 | 0.97 | 2.32 |
| UW | 100 | -0.11 | [-0.25;0.033] | 2.53 | 2.25 | 0.975 | 2.25 | |
| SEM | 1000 | -0.34 | [-0.35;-0.34] | 0.061 | 0.061 | 0 | 0.35 | 2.00 |
| LM | 1000 | -0.27 | [-0.29;-0.26] | 0.27 | 0.28 | 0.80 | 0.39 | 1.79 |
| UW | 1000 | 0.0037 | [-0.04;0.047] | 0.78 | 0.70 | 0.97 | 0.70 | |
| Simulation experiment 2 | | | | | | | | |
| | $n$ | Bias | 95% CI Bias | Estimated SE | Empirical SE | Coverage | MSE | MSE ratio |
| SEM | 100 | -0.28 | [-0.29;-0.27] | 0.18 | 0.18 | 0.63 | 0.33 | 5.48 |
| LM | 100 | -0.30 | [-0.33;-0.27] | 0.46 | 0.48 | 0.88 | 0.57 | 3.18 |
| UW | 100 | -0.083 | [-0.20;0.030] | 2.14 | 1.81 | 0.98 | 1.81 | |
| SEM | 1000 | -0.27 | [-0.28;-0.27] | 0.056 | 0.056 | 0.004 | 0.28 | 2.00 |
| LM | 1000 | -0.27 | [-0.28;-0.27] | 0.14 | 0.15 | 0.52 | 0.31 | 1.81 |
| UW | 1000 | 0.0032 | [-0.032;0.038] | 0.66 | 0.56 | 0.98 | 0.56 | |
| Simulation experiment 3 | | | | | | | | |
| | $n$ | Bias | 95% CI Bias | Estimated SE | Empirical SE | Coverage | MSE | MSE ratio |
| SEM | 100 | -1.00 | [-1.02;-0.98] | 0.29 | 0.30 | 0.070 | 1.044 | 0.39 |
| LM | 100 | -0.20 | [-0.22;-0.17] | 0.38 | 0.38 | 0.90 | 0.43 | 0.95 |
| UW | 100 | -0.0035 | [-0.030;0.023] | 1.20 | 0.41 | 1 | 0.41 | |
| SEM | 1000 | -0.996 | [-1.00;-0.99] | 0.091 | 0.092 | 0 | 1.00 | 0.12 |
| LM | 1000 | -0.196 | [-0.20;-0.19] | 0.12 | 0.12 | 0.62 | 0.23 | 0.52 |
| UW | 1000 | 0.0032 | [-0.0045;0.011] | 0.37 | 0.12 | 1 | 0.12 | |

Table 4.2: *Simulation results for the case where $X$ affects $L$. CI=Confidence interval, MSE=mean squared error, MSE ratio= ratio of the MSE of the unweighted estimator versus the MSE of the other estimators.*

| Simulation experiment 1 | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | $n$ | Bias | 95% CI Bias | Estimated SE | Empirical SE | Coverage | MSE | MSE ratio |
| SEM | 100 | -0.14 | [-0.15;-0.13] | 0.15 | 0.16 | 0.83 | 0.21 | 4.48 |
| LM | 100 | -0.047 | [-0.10;0.0089] | 0.88 | 0.90 | 0.94 | 0.90 | 1.04 |
| UW | 100 | -0.041 | [-0.99;0.017] | 1.35 | 0.94 | 0.99 | 0.94 | |
| SEM | 1000 | -0.14 | [-0.14;-0.13] | -.048 | 0.048 | 0.18 | 0.14 | 2.07 |
| LM | 1000 | -0.0019 | [-0.016;0.019] | 0.27 | 0.28 | 0.95 | 0.28 | 1.04 |
| UW | 1000 | 0.0026 | [-0.015;0.020] | 0.41 | 0.29 | 0.99 | 0.29 | |
| Simulation experiment 2 | | | | | | | | |
| | $n$ | Bias | 95% CI Bias | Estimated SE | Empirical SE | Coverage | MSE | MSE ratio |
| SEM | 100 | -0.072 | [-0.081;-0.063] | 0.14 | 0.15 | 0.90 | 0.17 | 3.12 |
| LM | 100 | -0.025 | [-0.054;0.042] | 0.46 | 0.47 | 0.93 | 0.47 | 1.13 |
| UW | 100 | -0.019 | [-0.052;0.014] | 1.38 | 0.53 | 1 | 0.53 | |
| SEM | 1000 | -0.068 | [-0.070;-0-065] | 0.045 | 0.046 | 0.68 | 0.081 | 1.85 |
| LM | 1000 | 0.0015 | [-0.0077;0.011] | 0.14 | 0.15 | 0.94 | 0.15 | 1.00 |
| UW | 1000 | 0.0022 | [-0.0072;0.012] | 0.42 | 0.15 | 1 | 0.15 | |
| Simulation experiment 3 | | | | | | | | |
| | $n$ | Bias | 95% CI Bias | Estimated SE | Empirical SE | Coverage | MSE | MSE ratio |
| SEM | 100 | | | | | | | |
| LM | 100 | 0.0012 | [-0.023;0.025] | 0.37 | 0.38 | 0.94 | 0.38 | 1.11 |
| UW | 100 | 0.0067 | [-0.0019;0.033] | 1.11 | 0.42 | 1 | 0.42 | |
| SEM | 1000 | | | | | | | |
| LM | 1000 | 0.0037 | [-0.0033;0.011] | 0.11 | 0.11 | 0.95 | 0.11 | 1.09 |
| UW | 1000 | 0.0039 | [-0.0034;0.011] | 0.34 | 0.12 | 1 | 0.12 | |

Table 4.3: *Simulation results in case $X$ does not affect $L$. CI=Confidence interval, MSE=mean squared error, MSE ratio= ratio of the MSE of the unweighted estimator versus the MSE of the other estimators.*

| with effect of $X$ on $L$ | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | $n$ | Bias | 95% CI Bias | Estimated SE | Empirical SE | Coverage | MSE | MSE ratio |
| SEM | 100 | -0.27 | [-0.29;-0.26] | 0.25 | 0.27 | 0.79 | 0.38 | 6.74 |
| LM | 100 | -0.30 | [-0.34;-0.26] | 0.66 | 0.68 | 0.92 | 0.75 | 3.41 |
| UW | 100 | -0.097 | [-0.26;0.06] | 2.80 | 2.56 | 0.98 | 2.56 | |
| SEM | 1000 | -0.27 | [-0.28;-0.27] | 0.079 | 0.084 | 0.069 | 0.29 | 2.62 |
| LM | 1000 | -0.28 | [-0.29;-0.26] | 0.20 | 0.20 | 0.72 | 0.34 | 2.24 |
| UW | 1000 | <0.001 | [-0.047;0.047] | 0.87 | 0.76 | 0.97 | 0.76 | |
| without effect of $X$ on $L$ | | | | | | | | |
| | $n$ | Bias | 95% CI Bias | Estimated SE | Empirical SE | Coverage | MSE | MSE ratio |
| SEM | 100 | -0.67 | [-0.081;-0.054] | 0.20 | 0.21 | 0.92 | 0.22 | 3.50 |
| LM | 100 | -0.023 | [-0.064;0.019] | 0.65 | 0.66 | 0.95 | 0.66 | 1.17 |
| UW | 100 | -0.023 | [-0.071;0.024] | 1.58 | 0.77 | 1 | 0.77 | |
| SEM | 1000 | -0.069 | [-0.073;-0.065] | 0.064 | 0.066 | 0.81 | 0.10 | 2.00 |
| LM | 1000 | <0.001 | [-0.012;0.012] | 0.20 | 0.20 | 0.96 | 0.20 | 1.00 |
| UW | 1000 | <0.001 | [-0.013;0.012] | 0.47 | 0.20 | 1 | 0.20 | |

Table 4.4: *Simulation results for Simulation experiment 4, in which $U$ has a quadratic effect on $Y$. CI=Confidence interval, MSE=mean squared error, MSE ratio= ratio of the MSE of the unweighted estimator versus the MSE of the other estimators.*

## 4.6 Brief overview of literature about estimation of direct effects

Methods for direct effects estimation have long been around, with extensive developments mainly within the structural equations literature (see e.g. Baron and Kenny, 1986; MacKinnon et al., 2002). These methods ignore the problem of confounding of the association between the intermediate variable and the outcome, and will therefore not be further discussed in this work. One of the first rigorous accounts of direct effects estimation, which acknowledges and clarifies the problems raised by such confounders, is Robins and Greenland (1992). These authors propose the G-computation algorithm to obtain controlled direct effect estimates adjusted for confounding bias. Later, Robins (1999b) discussed problems related to using the G-computation formula: (a) that is computationally complex; (b) that it is heavily model dependent; and (c) that the so-called null paradox guarantees rejection of the null hypothesis of no direct effect with probability approaching 1 as the sample size increases. The latter is due to the G-computation formula, like structural equation models, being based on models that do not carry direct effect parameters.

In view of this, Robins (1999b) proposes structural nested direct effect models which parameterize controlled direct effects (see Chapter 6). An estimation method based on inverse probability weighting is proposed, whereby each subject's data is inversely weighted by a conditional distribution of the intermediate variable. This method works well when the intermediate variable is categorical, but suffers from instability when this variable is absolutely continuous or when strong predictors of the intermediate variable exist (Vansteelandt, Goetgeluk et al., 2008; Goetgeluk et al., 2008). Van der Laan and Petersen (2004) develop similar methods based on marginal structural models for multiple interventions. Both methods have been proposed for longitudinal data.

Robins and Greenland (1992), Pearl (2001) and Petersen, Sinisi and van der Laan (2006) argue that natural direct effects are often more meaningful effect estimands than controlled direct effects. Robins and Greenland (1992) argue that these are extremely difficult to identify and suggest that

they are equivalent to controlled direct effects under the assumption of no interaction between the exposure and the intermediate variable. Pearl (2001) avoids the no-interaction assumption, which is testable, and proposes an untestable assumption under which natural direct effects can be identified. This is further relaxed in Petersen, Sinisi and van der Laan (2006) and van der Laan and Petersen (2004).

Ten Have et al. (2007) avoid the assumption of no unmeasured confounders for the association between the intermediate variable and the outcome, by using an instrumental variables approach (with the target exposure taken as the instrumental variable). Their method is of interest in the context of randomized experiments, but involves untestable no-interaction assumptions and tends to yield very inefficient estimates. Using a somewhat related estimation principle, Frangakis and Rubin (2002) identify principal stratification direct effects, additionally assuming the non-existence of certain principal strata.

In Chapter 5, we develop a simple, novel estimation method for controlled direct effects, which is easily implemented in statistical software. It is based on the no-unmeasured confounders assumption of Robins (1999b), but offers much more stable and accurate inferences, even in case the intermediate variable is continuous. We apply the methodology on a family-based association study to estimate the direct effect of certain SNPs in the IL10-gene on asthma, that is not mediated through body mass. This chapter was originally written as a stand alone article for geneticists. In Chapter 6, we extend estimation of direct effects based on Robins' structural nested direct effect models by developing doubly robust estimators. This means that the assumption on which these models are based is relaxed which makes them of use in realistic situations. These doubly robust estimators offer more stable and accurate inferences than the estimators obtained with the structural nested direct effect model. In addition, we find that the simple estimation method of Chapter 5, is a special estimator in the class of estimators developed in Chapter 6.

# Chapter 5

# A general principle for the identification of direct causal genetic pathways in association studies


### Summary

In genetic association studies, different complex phenotypes are often associated with the same marker. Such associations can be indicative of common genetic causes, non-genetic/environmental links between the traits, or the gene causing one of the traits which in turn causes the other trait. The presence of these multiple possible scenarios can obscure the true causative association and impede its identification. To identify the phenotype(s) with the causal genetic effects, statistical methods are needed to distinguish among these different possible origins of the associations. Herein, we propose a simple, general adjustment principle that can be incorporated into many standard genetic association tests to infer that a SNP has a direct causal influence on a given trait other than through the SNP's influence on another correlated phenotype. The proposed adjustment requires an estimate of the effect of the intermediate phenotype on the target trait, and thus requires measurements on all important common risk factors of both traits. Given such measurements, the adjustment

is straightforward to compute and thus particularly relevant for genome-wide association studies, pathway analyses, and association studies with an integrative genomic component. Using simulation studies, we show that standard association tests without the proposed adjustment can be biased in the presence of a non-causal link between marker and phenotypes. The simulations confirm our theoretical derivations that the proposed methodology is unbiased in such situations. Its achieved power levels are almost identical to those of standard methodology in situations where standard methods are valid. An application of the principle to three genome-wide association studies illustrates its practical importance.

[Original co-authors: S. Vansteelandt, Irwin Waldman, Helen Lyon, Eric E. Schadt, Scott T Weiss and Christoph Lange]

## 5.1 Introduction

It is well-established that the findings of genetic association studies can be confounded and biased by genetic and/or phenotypic heterogeneity which is not accounted for in statistical analyses. Consequently, much effort has been devoted to the development of statistical analytic techniques that minimize the impact of such effects, in particular population admixture and stratification (Pritchard and Rosenberg, 1999; Devlin and Roeder, 1999; Price et al., 2006; Epstein et al., 2007). Relatively little is known, however, about situations in which the same SNP is associated with multiple phenotypes which are themselves associated other than through the SNP of interest. Given such related phenotypes, a true association of the SNP with one phenotype can also induce an association with the other phenotype even in the absence of a direct independent effect of the SNP on that phenotype (Smoller et al., 2000; Robins et al., 2001). Such situations arise in genome-wide association studies (GWAS), in pathway analyses of candidate genes, and in association studies that incorporate genomic information.

There are three common contexts in which this problem arises. First, in both medical and psychiatric genetics, disorders of interest frequently

overlap and their symptoms covary, a phenomenon termed 'comorbidity'. In such cases, a SNP associated with the target disorder will often also be associated with the comorbid disorder(s). This raises the question of whether such associations reflect the independent influence of the SNP on the comorbid disorders or merely the influence of the SNP on the target disorder, which then in return stimulates the occurrence of comorbid disorders. Similar issues are encountered in the pathway analysis of complex diseases. For example, recent independent genome-scans for obesity and nicotine addiction revealed associations between SNPs in the FTO gene and both BMI and Smoking Quantity (SQ) (Frayling et al., 2007; Bierut et al., 2007). It is not obvious whether the observed associations here are attributable to direct genetic effects of FTO on both phenotypes, or whether links between nicotine addiction and obesity induce associations with each phenotype that are not a direct influence of the FTO gene. Second, medical and psychiatric geneticists often seek to explain the association between a SNP and a disorder using constructs that are considered to be more direct and proximal influences of the gene of interest. Such constructs, often referred to as endophenotypes or intermediate phenotypes, may represent immediate biological products of the gene, aspects of physiological or neurological function or various cognitive/neuropsychological functions. One index of the utility and validity of such endophenotypes is that they account for all or at least part of the SNP's influence on the disorder of interest. Thus, researchers are interested in testing whether the SNP influences the endophenotype in addition to the disorder and, conversely, whether the SNP shows any residual association with the disorder after accounting for its effects on the endophenotype. A third, conceptually different situation is encountered in genetic association studies that incorporate genomic data, e.g., expression profiles. Here, associations between the same marker and both traits, the expression profile and the disease phenotype of interest, can be especially informative. If one is able to conclude that the observed association between the marker and the disease phenotype is caused by the marker's effects on the expression profile, this increases the validity of an association finding with the disease phenotype.

A first common approach to test whether a SNP shows any residual

association with the target phenotype other than through its effects on a related phenotype, is to regress the target phenotype on the related phenotype and to use the corresponding residuals as the phenotype of interest in the association analysis. This approach is commonly recommended in GWAS of complex diseases, although with the different purpose of reducing the environmental variance of the phenotype, consequently increasing the statistical power of the association test. A second approach is to test whether the SNP is associated with the target phenotype after adjusting for the other phenotype. In this article, we argue, using causal diagrams, that both approaches are fallible. For the first approach, this is partly because by removing the association between both phenotypes (through taking residuals), one risks to remove also part of the effect of the SNP on the target phenotype. For the second approach this is because adjustment for phenotypes, or more generally covariates, is only valid under the assumption that the influencing factors/covariates are not associated with the marker locus (Rosenbaum, 1984; Cole and Hernan, 2002). Nonetheless, in the context of GWAS which interrogate the entire human genome, the assumption that covariates are not associated with a relevant marker locus is generally untenable. In order to identify the true genetic pathways underlying complex diseases, it will be crucial to distinguish whether an observed genetic association with a phenotype is attributable to its non-marker relation with another phenotype that is itself influenced by the marker, or whether the observed association represents the independent effects of the SNP, thus indicating a direct causal genetic relationship between the marker locus and the phenotype.

To address this problem, we propose a simple, general principle and method to adjust the phenotype of interest for its relation with another phenotype. We develop a computationally simple adjustment approach that is applicable to both binary and quantitative traits. The proposed adjustment requires an estimate of the effect of the intermediate phenotype on the target trait, and thus requires measurements on all important common risk factors of both traits. Given such measurements, the adjustment is straightforward to compute and can be incorporated into many standard genetic association tests, which then test for the direct genetic effect of

the marker on the phenotype of interest. It is important to note that if the adjustment is applied to account for a 'false-positive' association and the other phenotype is not actually associated with the marker locus, the adjustment remains valid.

Using simulation studies, we demonstrate that, association tests with covariate adjustments that are based on intuitive regression approaches can be biased and provide incorrect results whenever some of these covariates have a non-causal relation with the target phenotype. In contrast, our simulation studies verify that association tests using the proposed adjustment approach are valid when all common risk factors/confounders of the target phenotype and these covariates are correctly taken into account, and show reasonable robustness against moderate degrees of residual confounding. The simulation studies also demonstrate that the approach is suffciently powered to detect causal genetic effects for realistic sample sizes. Further, when the proposed adjustment is applied to account for a 'false positive' association, the adjusted association test remains unbiased and is only slightly less powerful compared to standard regression approaches. This suggests the proposed approach is generally applicable for the adjustment of phenotypes in genetic association studies.

As a real data example, we present results from a GWAS in the Framingham Heart Study that suggested an association between a SNP and two phenotypes, FEV1 and BMI which was detected using a standard regression-based covariate adjustment. In contrast, application of the proposed adjustment method suggests that the 'true' effect of this SNP originated from its association with BMI and that the observed association with FEV1 is attributable to this true association. This conclusion is then confirmed by replication replications in an independent GWAS, the CAMP study (Group CAMP, 1999).

## 5.2 Fallacies of intuitive regression adjustments

Suppose that in the study of interest, $n$ subjects have been genotyped at a specified marker locus and their coded genotype are denoted by $X_i, i = 1, ..., n$. If the selected sample is a family-based study, we further assume

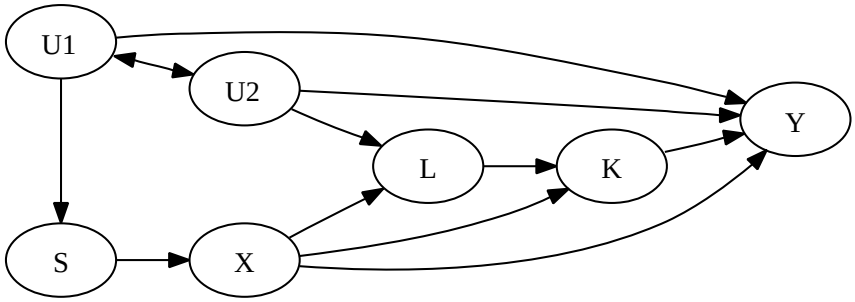


Figure 5.1: *Causal diagram illustrating the confounding of the genetic association between the primary phenotype $Y$ and the SNP $X$. The variable $K$ denotes the intermediate phenotype, $X$ the SNP, and $S$ a collection of measured factors (e.g., parental genotypes in the case of a family-based study) inducing population admixture. $U_1$ denotes a collection of unmeasured factors that allow for confounding due to population admixture and $U_2$ a collection of common risk factors of both phenotypes.*

that additional genotype data on other family-members are available so that the expected marker score, $E(X_i|S_i)$, can be computed conditional on Mendelian transmissions. When parental data are available, the variable $S_i$ denotes the parental genotypes; otherwise it represents the sufficient statistic by Rabinowitz and Laird (2000).

Two phenotypes, $K_i$ and $Y_i$, have been recorded for the $i$th subject. Both phenotypes are influenced by the shared, non-marker related set of factors, $L_i$, and are thus correlated. Assume that the first 'intermediate' phenotype $K_i$ has been tested for association with the SNP, and a significant association has been observed. Now, given the established association between the SNP and the phenotype $K_i$, our goal is to test for an association between the 'target' phenotype $Y_i$ and the SNP that cannot be explained by the existing genetic association with $K_i$ and the correlation between both phenotypes.

To understand the problems of standard intuitive regression approaches, we use causal diagrams (Pearl, 1995; Robins, 2001; Robins et al., 2001). These postulate the causal relationships between all measurements of in-

terest by means of directed arrows, as in Figure 5.1. Here, $S$ encodes the parental genotypes in the case of a family-based study, and measured factors inducing population admixture otherwise. Further, $U_1$ is a collection of unmeasured factors that allow for confounding due to population admixture and $U_2$ a collection of unmeasured common risk factors of $L$ and $Y$. For this diagram to be causal, it must satisfy the following two basic assumptions:

1. The absence of an arrow between any two variables $A$ and $B$ encodes the assumption that $A$ exercises no direct causal effect on $B$. Figure 5.1 thus postulates that $L$ cannot affect $Y$ other than by modifying $K$. We will relax this restriction later.

2. The diagram includes all variables that jointly affect any two variables in the diagram. Figure 5.1 thus postulates that all risk factors for the intermediate phenotype $K$, which are also associated with the target phenotype $Y$, have been measured and are contained in $L$. The figure allows, however, for the presence of unmeasured common risk factors of $L$ and $Y$, and of $S$ and $Y$.

The double-headed arrow between $U_1$ and $U_2$ allows for both variables to be associated (i.e., it allows for an unmeasured common cause). Note that a critical assumption, embedded in Figure 5.1, is that the target phenotype $Y$ is affected by the intermediate phenotype $K$ and not the other way around. We will elaborate on this restriction in the discussion section.

Two variables in a causal diagram may be statistically associated along all paths that have no converging arrows (i.e., along all unbroken sequences of edges between those two variables, disregarding the direction of the arrows, in which no two arrows point to each other) (Pearl, 1995; Robins, 2001; Robins et al., 2001). Specifically, the genotype $X$ and primary phenotype $Y$ in Figure 5.1 may be associated because of a direct genetic effect (i.e., along the path $X - Y$), because of an indirect genetic effect (i.e., along the path $X - K - Y$) and because of population admixture (i.e., along the path $X - S - U_1 - Y$). Importantly, note that when the target phenotype is not directly genetically affected, then (in the absence of

population admixture) no association can be detected between the genotype and the target phenotype, unless there is a causal link between both phenotypes (i.e., unless $K$ causally affects $Y$). Indeed, a non-causal relationship between both phenotypes does not induce an association between the genotype and the target phenotype because the converging arrows along the paths $X-K-L-U_2-Y$ and $X-L-U_2-Y$ transmit no association.

We will now use the causal diagram in Figure 5.1 to gain insight into the validity of common regression approaches to test the hypothesis whether the SNP directly affects the target phenotype other than through the intermediate phenotype $K$. A first common approach is to eliminate the effects of the established association with the other phenotype $K_i$ by regressing the target phenotype $Y_i$ on $K_i$ and using the residuals of $Y_i$ as the new phenotype in the association test. This approach has the disadvantage that the residuals remove the overall association between both phenotypes, which mixes the effect of the intermediate phenotype on the target trait (i.e., along the path $K-Y$), with spurious (i.e., non-causal) associations through the SNP $X$ (i.e., along the paths $K-X-Y$, $K-L-X-Y$ and $K-X-S-(U_1,U_2)-Y$), and by spurious association other than through the SNP $X$ (i.e., along the path $K-L-(U_1,U_2)-Y$). By not solely removing the causal effect of the intermediate phenotype on the target trait, the residuals may be associated with the SNP, even when it has no direct influence on the target trait. In particular, suppose that the SNP directly influences $K$, but not $Y$, and that $K$ has no effect on $Y$. Then the SNP has neither a direct, nor an indirect effect on $Y$. Nonetheless, the residual, say $Y-\gamma K$, will have $\gamma \neq 0$ because $Y$ is spuriously associated with $K$ along the path $K-L-U_2-Y$, and will be associated with the SNP by the fact that $K$ is influenced by it.

An alternative common approach to test the hypothesis that the SNP has a direct influence on the target phenotype other than through $K$, is to measure the association between $X$ and $Y$ conditional on $K$. To appreciate the impact of such adjustment for $K$, note that adjustment for a variable $K$ on a path between two variables $X$ and $Y$ in the causal diagram blocks the association between those variables along that path. This is true except when $K$ is a (descendant of a) collider along that path, in which case

an association is induced (Pearl, 1995; Robins, 2001; Robins et al., 2001). Stratification of the analysis on the intermediate phenotype $K$ thus removes the indirect effect of $X$ on $Y$, but at the same time induces a spurious association along the path $X-K-U_2-Y$ because $K$ is a collider along that path. Additional adjustment for the confounders $L$ blocks this assocation, but induces a new, non-causal association along the path $X-L-U_2-Y$.

In summary, traditional approaches for estimating/testing direct genetic effects, either based on an analysis of residuals or based on ordinary regression adjustment, may yield biased inferences whenever, as is likely the case, the association between the intermediate and target phenotype is confounded. Traditional regression adjustment for measured confounders remains problematic when (some of) these confounders are themselves influenced by the the target SNP. In the next section, we propose an alternative adjustment principle which is valid even when these confounders may be genetically affected.

## 5.3 A general principle to test for causal direct genetic effects

For simplicity, we illustrate the principle first for the association analysis of quantitative traits, either in population-based designs or in family-based designs. For both scenarios, i.e. population-based studies and family-based designs, we assume that the selected association test $T$ (e.g., a standard score test, Wald test or likelihood ratio test) has the general form

$$T = \sum_{i=1}^{n} T_i, \tag{5.1}$$

where $T_i$ denotes the contribution of the $i$th subject to the test statistic. To keep the notation simple and without loss of generality, we assume that the expected value of the test statistic $T$ is 0 (i.e. $E(T) = 0$) under the null-hypothesis of no association between the selected phenotype and the marker locus.

In order to derive a valid test for the direct association between the

target phenotype $Y$ and the SNP that is not influenced by the existing association between the intermediate phenotype $K$ and the same SNP, we need to estimate the non-marker related effect of $K_i$ on the target phenotype $Y_i$. Inferring this requires knowledge regarding all common risk factors on both phenotypes, as this is a prerequisite for disentangling a spurious association between both traits from a real effect. A test for a direct genetic effect of the SNP on the target phenotype $Y_i$ will therefore require an assessment of all shared risk factors influencing both phenotypes other than the SNP of interest. Nonetheless, as we will see in the simulation study this assumption can be relaxed and applies only to the major risk factors underlying both phenotypes, which are typically known.

For simplicity, we assume here that the target phenotype $Y_i$ and the intermediate phenotype $K_i$ share one common risk factor $L_i$. Then, in a population-based study, the following linear model can be used to assess how the phenotype $K_i$ influences the target phenotype $Y_i$

$$E(Y_i) = \gamma_0 + \gamma_1 K_i + \gamma_2 X_i + \gamma_3 L_i \tag{5.2}$$

In a family-based study, the expected marker-score, $E(X_i|S_i)$, would be added to the model to maintain robustness against population stratification, i.e.

$$E(Y_i) = \gamma_0 + \gamma_1 K_i + \gamma_2 X_i + \gamma_3 L_i + \gamma_4 E(X_i|S_i). \tag{5.3}$$

In both equations, $\gamma_0, \gamma_1, \gamma_2, \gamma_3$ and $\gamma_4$ denote the mean parameters. It is important to note that these models include the offspring genotype $X_i$ as well as the common risk factor $L_i$ of both phenotypes, in order to ensure that $\gamma_1$ represents the true effect of $K_i$ on $Y_i$ and no spurious association. This will guarantee later during the computation of the adjusted phenotypes (i.e., the residuals) that only the effect of $K_i$ is removed from the target phenotype, but a potential direct association between $Y_i$ and the target SNP is maintained. Indeed, it follows upon applying the principles outlined in the previous section to the causal diagram in Figure 5.1, that adjustment for $X_i, L_i$ and $S_i$ removes spurious associations between both phenotypes and thus reveals the causal effect of $K_i$ on $Y_i$.

Using ordinary least squares to estimate all parameters in model (5.2) or (5.3), the target phenotype $Y_i$ can be adjusted for just the effect that the phenotype $K_i$ has on the target phenotype $Y_i$,

$$\tilde{Y}_i = Y_i - \overline{y} - \hat{\gamma}_1(K_i - \overline{k}) \tag{5.4}$$

where $\hat{\gamma}_1$ is the ordinary least squares estimate for $\gamma_1$ in model (5.2) or (5.3) and $\overline{y}$ and $\overline{k}$ are the observed phenotypic means of $Y$ and $K$, respectively, in the sample. The phenotype adjustment (5.4) here deliberately only involves the other phenotype $K_i$ and not the shared risk factor $L_i$. Although this may seem counterintuitive and contrary to standard practice, including factors such as $L_i$ in the phenotypic adjustment would introduce bias to the extent that the common risk factor $L_i$ is itself associated with the target SNP. If the residuals $\tilde{Y}_i$ are computed based on the adjustment formula (5.4), this problem is avoided because the adjustment removes the effect of $K_i$ on $Y_i$ (i.e., the direct edge from $K_i$ into $Y_i$ on Figure 5.1) and thus also the indirect genetic effect (see Appendix 5.A1 for a more detailed argument).

Using the adjusted phenotype (i.e., the residual $\tilde{Y}_i$) as the target phenotype, we can construct standard association tests for quantitative traits in either population-based or family-based designs. For example, in a population-based setting, the adjusted phenotype $\tilde{Y}_i$ can be tested for association with a standard regression approach, i.e. each subject's contribution to the test statistic is given by

$$T_i = \{X_i - E(X_i)\}\,\tilde{Y}_i \tag{5.5}$$

For family-based studies, we can construct an association test based on a standard FBAT statistic (Laird et al., 2000) by defining the contribution of each offspring to be

$$T_i = \{X_i - E(X_i|S_i)\}\,\tilde{Y}_i \tag{5.6}$$

Both association tests, (5.5) and (5.6), will then test for a direct association between the SNP and the target phenotype $Y_i$ (other than through its association with the phenotype $K_i$). More generally, as we show in Appendix

5.A1, any association test statistic $T$ which is linear in the phenotype and which utilizes the adjusted phenotype $\tilde{Y}_i$ (equation (5.4)) as the target phenotype will provide a valid test for the null hypothesis that there is no direct effect of the SNP on the target phenotype $Y_i$, provided that model (5.2) or (5.3) is correctly specified and includes all common risk factors for both phenotypes, $Y$ and $K$. Under this condition, the expected value of the test statistic $T$ will be zero, $E(T) = 0$ when $T$ is computed based on the adjusted phenotype (equation (5.4)). Further, for family-based tests, we show in Appendix 5.A1 that, under the null hypothesis of no direct effect, the modified FBAT-statistic (5.6) remains robust against confounding due to population admixture, so long as the population admixture on both phenotypes is w.r.t. unrelated causes (since otherwise there could exist unmeasured common causes of both phenotypes).

The proposed phenotype adjustment (equation (5.4)) of the association test $T$ for a direct genetic effect includes a parameter estimate for $\gamma_1$ that is obtained by fitting model (5.2) or (5.3). Given that the imprecision of the estimate for $\gamma_1$ must be acknowledged in the computation of the asymptotic variance of the test statistic, the standard variance of the selected association test is no longer applicable.

In Appendix 5.A1, we show that the standardized association test statistic for the adjusted phenotype $T^2/(n\Sigma)$ follows a chi-square distribution with 1 degree of freedom under the null hypothesis of no direct effect, where the variance of the test statistic, $\Sigma$, is given by

$$\begin{aligned} \Sigma &= Var(\tilde{T}_i) \\ \text{with} \quad \tilde{T}_i &= T_i(\tilde{Y}_i) - E\left[T_i'(\tilde{Y}_i)K_i\right]\frac{\left(K_i - \mu_K^{(i)}\right)}{\sigma_K^2}\epsilon_i \end{aligned}$$

where $T_i(P)$ denotes the contribution of the $i$th subject to the association test statistic for the target phenotype $P$ and $T_i'(P)$ the first order derivative of $T_i(P)$ w.r.t. $P$ (e.g., for population-based tests, we have $T_i'(\tilde{Y}_i) = X_i$ in (5.5) and, for family-based tests, $T_i'(\tilde{Y}_i) = X_i - E(X_i|S_i)$ in (5.6)). The variable $\epsilon_i$ is the residual in model (5.2). In population-based designs, the parameters $\mu_K$ and $\sigma_K^2$ are obtained by fitting a linear regression for $K_i$

with the covariates $L_i$ and $X_i$. For family-based studies, the covariate $E(X\,|S_i\,)$ has to be included as well. The predicted value for $K_i$ is then defined by $\mu_K^{(i)} = E(K\,|L_i, X_i\,)$ or by $\mu_K^{(i)} = E(K\,|L_i, X_i, E(X\,|S_i\,)\,)$. The residual variance in the model is denoted by $\sigma_K^2$.

When the phenotype of interest $Y_i$ is dichotomous, e.g. affection status, the proposed adjustment can be extended provided that a relative risk model and a log-link function is assumed. The technical details are discussed in Appendix 5.A1.

## 5.4 Data analysis: An application to the Framingham Heart Study, the British Birth Cohort and the CAMP study

We evaluated the practical relevance of the proposed adjustment principle by an application to 3 genome-wide association studies: a 100K Affymetrix scan in the family-plates of the Framingham Heart Study (1,400 probands) (Herbert et al., 2006), a 550K Illumina scan in the British Birth Cohort (genotype data on 1,430 probands,http://www.b58cgene.sgul.ac.uk/) and a 550K Illumina scan in 440 trios of the CAMP study (Group CAMP, 1999). As target phenotype, we selected the lung-function measurement FEV1, which was available in all 3 studies. Since the 3 studies were genotyped on different platforms (the Framingham Heart Study on 100K Affymetrix, the British Birth Cohort and CAMP on Illumina 550K), we selected the 32,121 SNPs for the analysis that are common among both platforms.

As the first step, we analyzed FEV1 at exam 1 in the family-plates of the Framingham Heart Study. Using a standard regression approach, we adjusted FEV1 for height, height$^2$, gender, weight and age, and then used the residuals as the target phenotype in the analysis. All statistical analysis was conducted under an additive mode of inheritance. Since the Framingham Heart Study is a family-based study, we applied the weighted Bonferroni-testing strategy by Ionita-Laza et al. (2007). the testing strategy evaluates the evidence for association at a population-level and then

estimates the conditional power of the FBAT-statistic for each marker in the first step. In the second step of the testing strategy, FBAT-statistics are computed for all markers. Their significance is assessed based on individually adjusted $\alpha$-levels that maintain the overall type-1 error and that are weighted based on the conditional power estimate for the corresponding marker.

When the weighted Bonferroni-approach was applied to the 32,121 SNPs that are on both genotyping platforms, none of the SNPs reached genome-wide significance. However, the SNP (rs2415815) with the highest conditional power estimate had an unadjusted FBAT p-value of 0.0234, warranting additional analysis. When the covariates were tested for association with rs2415815, an association with weight was observed (p-value of 0.0054 for FBAT adjusted for age and gender). Both associations between the SNP and the two phenotypes were then verified in the CAMP study and the British Birth Cohort (Table 5.1). For weight, SNP rs2415815 has nominal significant p-values in CAMP and the British Birth Cohort (measured as BMI), while the association tests with FEV1 are not significant in either studies. Given the established link between asthma and obesity for which FEV1 and weight (Olivetti et al., 2006; Yuan et al., 2002; Gessner and Chimonas, 2007; Sin et al., 2004; Demissie et al., 1998; Tavernas et al., 2006; Camargo et al., 1999) are endophenotypes, the inconsistent replications of FEV1 in CAMP and FHS were re-analyzed with the proposed adjustment procedure. For the British Birth Cohort, we did not have access to the raw data and the proposed adjustment could not be computed.

The lung-function measurement FEV1 was adjusted for its covariates, for the SNP and for weight. The phenotype weight was adjusted for gender, age and height, for the SNP and for FEV1. Based on the adjusted phenotype, the FBAT-statistics in CAMP and FHS were re-calculated (Table 5.1). The associations with weight remained significant after the adjustment for a potential genetic association with FEV1. However, when we accounted for a potential genetic association between the SNP and weight, the association tests for FEV1 in FHS and in CAMP were no longer significant (Table 5.1). Our results suggest that the originally observed association with FEV1 in the FHS may have been attributable to the association with

| Phenotype | Adjustment | FHS | CAMP | BBC |
|---|---|---|---|---|
| FEV1 | standard adjustment | 0.0376 | 0.5920 | 0.436 |
| | proposed adjustment | 0.1012 | 0.7790 | NA |
| Weight | standard adjustment | 0.0054 | 0.00053 | 0.0445* |
| | proposed adjustment | 0.0088 | 0.0033 | NA |

Table 5.1: *Association with rs2415815 in the Framingham Heart Study (FHS), the CAMP Study and the British Birth Cohort (BBC). * measured as BMI.*

weight and that there is no evidence of a direct genetic effect of rs2415815 on FEV1 other than through weight.

## 5.5 Simulation Study

Using simulation studies, we assess the type-1 error, the power and the robustness of the new approach and compare it to the two standard approaches earlier described. The new principle is evaluated under various conditions, including scenarios in which there are unmeasured risk factors that are common for both phenotypes and that are not included in the adjustment formula. In all simulations, we focus on quantitative traits and assume that there is no ascertainment condition. A sample size of 1,000 probands is selected. All simulation results that are presented in this chapter are based on 5,000 replicates. With the data analysis example in the Framingham Heart Study in mind, the phenotype of interest $Y$ is simulated so that it resembles the FEV1 phenotype in that application. The second phenotype $K$ is weight and the set of common confounding variables is given by height and age, which are denoted by $L$ and $S$, respectively. First the genotype data is generated by drawing from a Binomial distribution with the specified marker allele frequency. Using the genotype data, all phenotypic variables are simulated from normal distributions under the causal diagrams of Figure 5.2, with phenotypic means and variances that were observed in the application to the Framingham Heart Study. In addition,

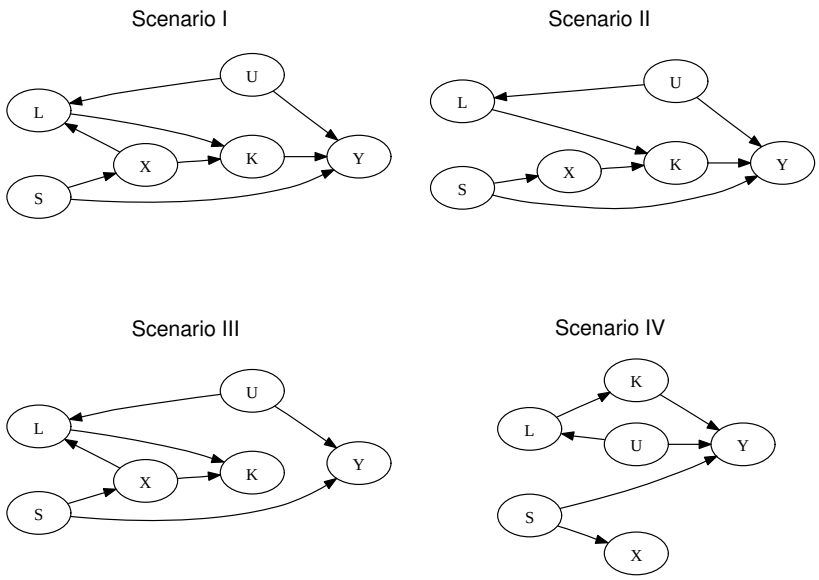


Figure 5.2: Causal diagrams illustrating the data generating mechanism under simulation scenarios I-IV.

unless otherwise specified, effect sizes were also chosen to match the data application.

Five tests are evaluated each time. The first two test the association between the target phenotype $Y$ and the marker locus with standard Wald tests that are applied to the adjusted phenotypic residuals, where the residuals are obtained from linear regressions adjusting for either $(S, K)$ or $(S, K, L)$. The next two test this association with standard Wald tests that are applied to the phenotype itself, adjusting for either $(S, K)$ or $(S, K, L)$. Finally, we evaluate the proposed adjustment principle.

### Empirical significance level

The first set of simulation studies is conducted under the null hypothesis of no direct genetic effect on the target phenotype $Y$. In order to assess

the robustness of the adjustment principle against spurious associations, we consider the following scenarios corresponding to the causal diagrams in Figure 5.2:

- In the first scenario, we assume that there is a direct genetic effect of the marker on the intermediate phenotype $K$ and on the common covariate $L$. Each genetic effect has a locus specific heritability of 1%. The intermediate phenotype $K$ explains 1% of the phenotypic variation in $Y$, creating a spurious association between the SNP and $Y$. Under these conditions, a standard adjustment for the covariates $(L, Z)$ will provide correct results for the Wald test.

- In the second simulation experiment, the first scenario is modified so that there is no genetic effect on the confounder $L$, but the genetic association with the intermediate phenotype $K$ is still present.

- The third simulation experiment varies from the first scenario with respect to the association between the phenotypes $Y$ and $K$. While the common confounder $L$ is still associated with the marker locus in the third scenario, there is no link anymore between the intermediate phenotype $K$ and the target phenotype $Y$, making an adjustment for the intermediate phenotype $K$ in the association analysis unnecessary.

- In the fourth simulation experiment, the second scenario is modified so that there is also no genetic effect on the intermediate phenotype $K$.

For a variety of allele frequencies, the estimated nominal significance levels are shown in Table 5.2 for the five different tests (performed at the 5% significance level).

The Wald test based on the new adjustment principle maintains the specified significance level well in all four scenarios and throughout the entire range of allele frequencies. This is true regardless of whether the intermediate phenotype affects the target phenotype $Y$ or not, and of whether the confounders $L$ for the association between both phenotypes are genetically affected. However, Wald tests that are based on the standard

adjustments for $(S, K, L)$ or $(S, K)$ generally fail to preserve the theoretical $\alpha$-level, with the exception of Scenarios II and IV. In the latter scenarios, the Wald test based on the standard adjustment for the confounding variables $(S, K, L)$ maintains the significance level because $L$ is not genetically affected in these scenarios. In Scenario IV, Wald tests that are based on residuals, adjusting for $(S, K, L)$ or $(S, K)$, additionally maintain the significance level because $Y$ and $K$, and thus the residuals, are jointly independent of the genotype in this scenario, after adjustment for $S$. While, as in these two scenarios, there are instances in which standard adjustments provides correct $\alpha$-levels for the Wald tests, this can only be achieved if the underlying genetic architecture is known and all necessary confounding variables are included in the standard standard adjustment. The proposed adjustment principle does not require any prior knowledge about potential links and genetic associations between the phenotypes and the covariates, and maintains the significance level in all considered scenarios.

| | | Wald test on residuals adjusted for | | Wald test on trait adjusted for | | Proposed adj. principle |
|---|---|---|---|---|---|---|
| S | Freq | (S,K,L) | (S,K) | (S,K,L) | (S,K) | (D) |
| 1 | 0.05 | 0.080 | 0.150 | 0.044 | 0.134 | 0.051 |
| | 0.1 | 0.085 | 0.079 | 0.065 | 0.075 | 0.047 |
| | 0.15 | 0.092 | 0.059 | 0.079 | 0.057 | 0.053 |
| | 0.2 | 0.083 | 0.047 | 0.075 | 0.047 | 0.047 |
| | 0.25 | 0.089 | 0.057 | 0.081 | 0.056 | 0.056 |
| | 0.3 | 0.093 | 0.058 | 0.086 | 0.058 | 0.057 |
| | 0.35 | 0.083 | 0.051 | 0.075 | 0.051 | 0.047 |
| | 0.4 | 0.085 | 0.055 | 0.078 | 0.055 | 0.052 |
| | 0.45 | 0.095 | 0.056 | 0.086 | 0.056 | 0.053 |

| | | Wald test on residuals adjusted for | | Wald test on trait adjusted for | | Proposed adj. principle |
|---|---|---|---|---|---|---|
| S | Freq | (S,K,L) | (S,K) | (S,K,L) | (S,K) | (D) |
| 2 | 0.05 | 0.054 | 0.052 | 0.054 | 0.052 | 0.053 |
| | 0.1 | 0.045 | 0.045 | 0.045 | 0.045 | 0.048 |
| | 0.15 | 0.047 | 0.045 | 0.047 | 0.045 | 0.047 |
| | 0.2 | 0.048 | 0.049 | 0.048 | 0.049 | 0.049 |
| | 0.25 | 0.049 | 0.050 | 0.049 | 0.050 | 0.048 |
| | 0.3 | 0.049 | 0.048 | 0.049 | 0.048 | 0.049 |
| | 0.35 | 0.054 | 0.052 | 0.054 | 0.052 | 0.054 |
| | 0.4 | 0.052 | 0.050 | 0.052 | 0.050 | 0.052 |
| | 0.45 | 0.052 | 0.051 | 0.052 | 0.051 | 0.054 |
| 3 | 0.05 | 0.085 | 0.144 | 0.050 | 0.131 | 0.049 |
| | 0.1 | 0.083 | 0.077 | 0.065 | 0.074 | 0.046 |
| | 0.15 | 0.083 | 0.061 | 0.072 | 0.060 | 0.050 |
| | 0.2 | 0.090 | 0.051 | 0.083 | 0.051 | 0.050 |
| | 0.25 | 0.085 | 0.048 | 0.078 | 0.048 | 0.046 |
| | 0.3 | 0.081 | 0.050 | 0.075 | 0.050 | 0.049 |
| | 0.35 | 0.085 | 0.054 | 0.076 | 0.054 | 0.052 |
| | 0.4 | 0.083 | 0.051 | 0.075 | 0.050 | 0.048 |
| | 0.45 | 0.095 | 0.053 | 0.086 | 0.052 | 0.050 |
| 4 | 0.05 | 0.049 | 0.046 | 0.049 | 0.046 | 0.047 |
| | 0.1 | 0.051 | 0.053 | 0.051 | 0.053 | 0.052 |
| | 0.15 | 0.049 | 0.049 | 0.050 | 0.049 | 0.051 |
| | 0.2 | 0.052 | 0.053 | 0.052 | 0.053 | 0.050 |
| | 0.25 | 0.049 | 0.047 | 0.049 | 0.047 | 0.048 |
| | 0.3 | 0.052 | 0.050 | 0.052 | 0.049 | 0.052 |
| | 0.35 | 0.049 | 0.054 | 0.050 | 0.054 | 0.054 |
| | 0.4 | 0.047 | 0.047 | 0.047 | 0.047 | 0.049 |
| | 0.45 | 0.052 | 0.052 | 0.052 | 0.052 | 0.053 |

Table 5.2: *Empirical Type I errors at 5% significance level of Wald tests for genetic effects, (a) based on residuals adjusted for $(S, K, L)$, $(S, L)$, (b) directly adjusted for $(S, K, L)$, $(S, L)$, and (c) adjusted with the proposed adjustment principle (D). S=Scenario*

### Estimated statistical power

To assess whether the Wald tests based on the new adjustment principle have sufficient power to detect genetic effects of realistic magnitudes, we repeat the simulation study under the assumption that, in all the four scenarios, there is a direct genetic effect of the marker locus on the target phenotype $Y$. The locus-specific heritability of the genetic effect is specified to be 0.33%. The estimated power levels for the Wald tests based on the proposed adjustment principle are displayed in Table 5.3. In general, we find high power levels in all simulation experiments, except, as usual, at low allele frequencies ($< 0.10$).

Comparison of the attained power levels of the new adjustment principle is especially relevant vis-a-vis the $(S, K, L)$-adjustment in the second Scenario and vis-a-vis all other tests in the fourth Scenario. In these scenarios, these other approaches are also valid and are expected to yield higher power levels by the fact that they involve stronger assumptions. Interestingly, these standard approaches provides power levels that are essentially identical to those of the proposed adjustment principle. Since these scenarios are essentially the best case scenarios for the standard approaches, these results illustrate the potential of the proposed adjustment principle as a generally applicable tool in genetic association studies.

| | | Wald test on residuals adjusted for | | Wald test on trait adjusted for | | Proposed adj. principle |
|---|---|---|---|---|---|---|
| S | Freq | (S,K,L) | (S,K) | (S,K,L) | (S,K) | (D) |
| 1 | 0.05 | 0.520 | 0.119 | 0.406 | 0.106 | 0.341 |
| | 0.1 | 0.787 | 0.437 | 0.743 | 0.427 | 0.624 |
| | 0.15 | 0.904 | 0.709 | 0.891 | 0.706 | 0.787 |
| | 0.2 | 0.955 | 0.857 | 0.951 | 0.856 | 0.868 |
| | 0.25 | 0.971 | 0.906 | 0.969 | 0.906 | 0.919 |
| | 0.3 | 0.980 | 0.930 | 0.977 | 0.929 | 0.943 |
| | 0.35 | 0.988 | 0.943 | 0.986 | 0.943 | 0.958 |
| | 0.4 | 0.992 | 0.957 | 0.991 | 0.957 | 0.969 |
| | 0.45 | 0.993 | 0.962 | 0.992 | 0.962 | 0.975 |

| | | Wald test on residuals adjusted for | | Wald test on trait adjusted for | | Proposed adj. principle |
|---|---|---|---|---|---|---|
| S | Freq | (S,K,L) | (S,K) | (S,K,L) | (S,K) | (D) |
| 2 | 0.05 | 0.405 | 0.401 | 0.405 | 0.401 | 0.390 |
| | 0.1 | 0.655 | 0.655 | 0.655 | 0.655 | 0.642 |
| | 0.15 | 0.805 | 0.802 | 0.804 | 0.801 | 0.791 |
| | 0.2 | 0.883 | 0.878 | 0.883 | 0.878 | 0.874 |
| | 0.25 | 0.929 | 0.927 | 0.929 | 0.927 | 0.923 |
| | 0.3 | 0.947 | 0.945 | 0.947 | 0.945 | 0.944 |
| | 0.35 | 0.964 | 0.962 | 0.964 | 0.962 | 0.958 |
| | 0.4 | 0.973 | 0.970 | 0.973 | 0.970 | 0.967 |
| | 0.45 | 0.976 | 0.974 | 0.977 | 0.974 | 0.974 |
| 3 | 0.05 | 0.519 | 0.114 | 0.396 | 0.101 | 0.337 |
| | 0.1 | 0.787 | 0.442 | 0.742 | 0.434 | 0.618 |
| | 0.15 | 0.897 | 0.691 | 0.884 | 0.689 | 0.777 |
| | 0.2 | 0.959 | 0.859 | 0.955 | 0.859 | 0.873 |
| | 0.25 | 0.975 | 0.910 | 0.973 | 0.909 | 0.927 |
| | 0.3 | 0.984 | 0.927 | 0.983 | 0.927 | 0.947 |
| | 0.35 | 0.986 | 0.945 | 0.985 | 0.945 | 0.958 |
| | 0.4 | 0.991 | 0.959 | 0.991 | 0.959 | 0.972 |
| | 0.45 | 0.995 | 0.964 | 0.994 | 0.964 | 0.974 |
| 4 | 0.05 | 0.407 | 0.400 | 0.407 | 0.400 | 0.396 |
| | 0.1 | 0.647 | 0.643 | 0.647 | 0.643 | 0.636 |
| | 0.15 | 0.803 | 0.802 | 0.804 | 0.803 | 0.794 |
| | 0.2 | 0.887 | 0.882 | 0.887 | 0.882 | 0.875 |
| | 0.25 | 0.927 | 0.926 | 0.927 | 0.926 | 0.919 |
| | 0.3 | 0.953 | 0.950 | 0.953 | 0.950 | 0.947 |
| | 0.35 | 0.962 | 0.961 | 0.962 | 0.961 | 0.959 |
| | 0.4 | 0.968 | 0.967 | 0.968 | 0.967 | 0.966 |
| | 0.45 | 0.977 | 0.976 | 0.977 | 0.976 | 0.974 |

Table 5.3: *Empirical power at 5% significance level of Wald tests for genetic effects, (a) based on residuals adjusted for $(S, K, L)$, $(S, L)$, (b) directly adjusted for $(S, K, L)$, $(S, L)$, and (c) adjusted with the proposed adjustment principle (D). S=Scenario*

### Robustness of the adjustment principle when common confounding variables of both phenotypes are not included in the adjustment principle

The final series of simulation experiments is aimed to evaluate the robustness of the proposed approach against the omission of common confounders for both phenotypes. We therefore introduce a (normally distributed) non-genetic risk factor $U$ to explain 1% and 10%, respectively, of the phenotypic variation in both phenotypes. The variable $U^*$ will be considered unmeasured in the analysis. In the presence of such a variable, the proposed adjustment approach, like the standard approaches, will be biased because it will estimate the effect of the intermediate phenotype on the target phenotype without incorporating the common, unmeasured risk factor.

For a variety of allele frequencies, Table 5.4 shows the estimated nominal significance levels for Wald tests that are based on the proposed adjustment principle. Although, as predicted by our theoretical considerations, our approach no longer maintains the specified significance level, the impact appears negligible for applications. For an unknown confounding variable that explains 1% of the phenotypic variation ($r^2 = 0.01$), no observable departure from the theoretical 5%-level can be detected. When a confounding variables that explains 10% of the phenotypic variation in both phenotypes is omitted in the adjustment principle, the theoretical significance level is not maintained for small allele frequencies ($< 15\%$). Given the current epidemiologic knowledge and understanding of the phenotypes studied in genetic association study, the omission of common confounding variable with $r^2$ of 10% is extremely unlikely. Even for a $r^2$-range of about 1%, most confounding variables for phenotypes of complex disease are typically known.

In summary, our simulation studies suggest that the proposed adjustment principle performs well under realistic conditions. The approach maintains the significance level in the presence of spurious associations. For realistic genetic effect sizes, the approach achieves sufficient power, even in situations in which an adjustment is not required and standard approaches are optimal. The disadvantage of the approach, the required knowledge

| | | Wald test on residuals adjusted for | | Wald test on trait adjusted for | | Proposed adj. principle |
|---|---|---|---|---|---|---|
| $U^*$ | Freq | (S,K,L) | (S,K) | (S,K,L) | (S,K) | (D) |
| 1% | 0.05 | 0.061 | 0.172 | 0.029 | 0.155 | 0.052 |
| | 0.1 | 0.061 | 0.093 | 0.046 | 0.090 | 0.050 |
| | 0.15 | 0.071 | 0.065 | 0.059 | 0.065 | 0.051 |
| | 0.2 | 0.076 | 0.049 | 0.070 | 0.049 | 0.051 |
| | 0.25 | 0.075 | 0.053 | 0.069 | 0.052 | 0.052 |
| | 0.3 | 0.070 | 0.051 | 0.064 | 0.050 | 0.046 |
| | 0.35 | 0.078 | 0.055 | 0.069 | 0.054 | 0.053 |
| | 0.4 | 0.067 | 0.048 | 0.060 | 0.047 | 0.046 |
| | 0.45 | 0.069 | 0.059 | 0.063 | 0.058 | 0.054 |
| 10% | 0.05 | 0.057 | 0.228 | 0.028 | 0.207 | 0.099 |
| | 0.1 | 0.054 | 0.117 | 0.037 | 0.113 | 0.069 |
| | 0.15 | 0.057 | 0.072 | 0.048 | 0.070 | 0.057 |
| | 0.2 | 0.057 | 0.050 | 0.054 | 0.050 | 0.050 |
| | 0.25 | 0.058 | 0.050 | 0.053 | 0.050 | 0.048 |
| | 0.3 | 0.056 | 0.054 | 0.051 | 0.054 | 0.052 |
| | 0.35 | 0.052 | 0.050 | 0.046 | 0.050 | 0.048 |
| | 0.4 | 0.059 | 0.055 | 0.052 | 0.055 | 0.050 |
| | 0.45 | 0.056 | 0.059 | 0.050 | 0.058 | 0.054 |

Table 5.4: *Empirical Type I errors at 5% significance level of Wald tests for genetic effects, (a) based on residuals adjusted for* $(S, K, L)$, $(S, L)$, *(b) directly adjusted for* $(S, K, L)$, $(S, L)$, *and (c) adjusted with the proposed adjustment principle (D), in the presence of unmeasured confounding.*

of all common confounding variables, turns out to be of lesser concern in applications where the degree of unmeasured confounding is anticipated to be weak.

## 5.6 Discussion

In order to understand the genetic architecture of complex diseases, it is important to gain insight into the multifaceted relationships between complex phenotypes and their genetic associations. The origin of an observed genetic association between a SNP and the target phenotype can be attributable to a direct genetic effect or can be caused by a non-genetic link with another phenotype that is itself influenced by the marker locus of interest. In order to prioritize the follow-up of large-scale association studies in terms of replication strategies in other populations or, even more importantly, in terms of functional work, it is crucial to be able to distinguish between these different sources of genetic association.

In this chapter, we proposed an adjustment that can be incorporated into many genetic association tests and, thereby, enables the test to assess whether an observed association is caused by a genetic association with another phenotype, or whether it is attributable to a direct genetic effect. The approach is computationally simple and can be easily be implemented in most software packages. If the principle is applied to correct for a false positive association, the adjusted test remains valid and its power is decreased only marginally compared to standard adjustment. These properties make the proposed procedure a universally applicable adjustment principle in genetic association studies.

# Appendix 5.A1: Distribution of the test statistic

The proposed adjustment principle forms a special case of the 'unweighted estimator' in Goetgeluk, Vansteelandt and Goetghebeur (2008), which allows for more general complexities, such as arbitrary non-linear models for the expected outcome and gene-environment interactions between the genotype and intermediate phenotype. Below, we demonstrate the validity of this principle for the setting that we have considered in this article.

By using the adjusted phenotype $\tilde{Y}$ in the test statistic (5.5) or (5.6), we remove the arrow from $K$ to $Y$, and thus the indirect effect of $X$ on $Y$. This can be seen using the principles of causal diagrams (see Section 5.2) and explains intuitively why a standard association test, using the adjusted phenotype $\tilde{Y}$, is valid for testing direct genetic effects. More formally, suppose that the null hypothesis is true that $X$ has no effect on $Y$ other than through $K$. Let

$$E(Y|X, K, U_1, U_2) = \Phi\{\omega(U_1, U_2) + \gamma_1 K\} \tag{5.7}$$

where $\Phi$ is the identity link ($\Phi(x) = x$) or the exponential link ($\Phi(x) = \exp(x)$) and where $\omega(U_1, U_2)$ is an arbitrary function. This model does not involve $X$ because we are working under the null hypothesis of no direct effect. Furthermore, the parameter $\gamma_1$ in this model is the same as in model

$$E(Y|X, K, L, S) = \Phi\{\omega^*(X, L, S) + \gamma_1 K\}$$

(cfr. model (5.2)), which can be seen by inferring this model from model (5.7) upon noting that $Y \amalg (L,S)|K,X,U_1,U_2$ and $(U_1,U_2) \amalg K|L,X,S$ under the diagram of Figure 5.1, where $A \amalg B|C$ for variables $A, B$ and $C$ means that $A$ is conditionally independent of $B$, given $C$. It now follows that the test statistic (5.6) with $\tilde{Y} = Y - \gamma K$ when $\Phi$ is the identity link and $\tilde{Y} = Y \exp(-\gamma K)$ when $\Phi$ is the exponential link (and likewise the test statistic (5.5)) has mean zero at the null hypothesis. For the test statistic (5.6), for instance, this is because

$$E\left[\{X - E(X|S)\}\,\tilde{Y}\right] \;=\; E\left[\{X - E(X|S)\}\,\Phi\{\omega(U_1,U_2)\}\right]$$

and by the fact that $(U_1,U_2) \amalg X|S$ (which follows because the genotype $X$ is only directly affected by $S$).

When there is population admixture on both phenotypes w.r.t. unrelated causes (i.e., when the diagram of Figure 5.1 includes an additional unmeasured variable $U_3$ which simultaneously affects $S$ and $K$), then the proposed adjustment principle remains valid. This is because the principles of causal diagrams show that adjustment for $S, X$ and $L$ is then still sufficient to estimate the causal effect of $K$ on $Y$, where the adjustment for $S$ in model (5.3) happens through the adjustment for $E(X|S)$.

*Remark.* Note that Figure 5.1 implicitly assumes that common risk factors $L$ of both phenotypes do not themselves affect the target phenotype. Our test remains valid without this assumption. In that case, it tests for an association between the target SNP and target phenotype $Y$, other than through the given intermediate phenotype $K$.

We now derive the distribution of the test statistic (5.6). The derivation is analogous for the test statistic (5.5) upon substituting $X_i - E(X_i|S_i)$ with $X_i - E(X_i)$. Using M-estimation arguments (van der Vaart, 1998, p.48-60), the test statistic (5.6) may be adjusted for estimating $\gamma_1$ by calculating the adjusted test statistic

$$\tilde{T}_i \;\equiv\; \{X_i - E(X_i|S_i)\}\,\tilde{Y}_i - \lambda\,\{K_i - E(K_i|L_i,X_i,S_i)\}\,\epsilon_i \qquad (5.8)$$

which guarantees that the test statistic (5.8) is uncorrelated with the scores needed for estimation of $\gamma_1$. Under model (5.2), $\epsilon_i$ is the residual from that

model,

$$\lambda \equiv \frac{E\left[\{X_i - E(X_i|S_i)\}\, K_i\right]}{Var\left(K_i|L_i, X_i, S_i\right)}$$

and $E(K_i|L_i, X_i, S_i)$ is the fitted value from a linear regression of $K_i$ on $(L_i, X_i, S_i)$. For binary traits or counts obeying the multiplicative model

$$E(Y_i) = \exp(\gamma_0 + \gamma_1 K_i + \gamma_2 X_i + \gamma_3 S_i + \gamma_4 L_i) \tag{5.9}$$

the adjusted phenotype can be computed as

$$\tilde{Y}_i \equiv Y_i \exp(-\hat{\gamma}_1 K_i) - \mu$$

where $\hat{\gamma}_1$ now denotes the maximum likelihood estimate obtained by fitting the Poisson regression model (5.9). In that case, $\epsilon_i$ is the residual from that model,

$$\lambda \equiv \frac{E\left[\{X_i - E(X_i|S_i)\}\, \tilde{Y}_i K_i\right]}{Var\left(\mu_i^{1/2} K_i|L_i, X_i, S_i\right)}$$

$\mu_i$ is the fitted value under model (5.9) and $E(K_i|L_i, X_i, S_i)$ is the fitted value from a weighted linear regression of $K_i$ on $(L_i, X_i, S_i)$, with weights $\mu_i$.

By the Central Limit Theorem, $n^{-1/2} \sum_{i=1}^{n} \tilde{T}_i$ has a normal distribution in large samples with mean zero at the null hypothesis and variance $\Sigma$ which can be estimated by the sample variance of $\tilde{T}_i$. Squaring and noting that $\sum_{i=1}^{n} \{K_i - E(K_i|L_i, X_i, S_i)\}\, \epsilon_i = 0$ by construction, yields the distribution of the test statistic as reported in Section 5.2.

# Chapter 6

# Estimation of controlled direct effects


## Summary

When regression models adjust for mediators on the causal path from exposure to outcome, the regression coefficient of exposure is commonly viewed as a measure of the direct effect of exposure. The term 'direct effect' then indicates the total effect of exposure on outcome, minus the effect that is due to an exposure effect on the mediator. This interpretation can be misleading, even with randomly assigned exposure. This happens because adjustment for post-exposure measurements introduces bias whenever their association with outcome is confounded by more than just the exposure. By the same token, additional adjustment for such confounders stays problematic when these confounders are themselves affected by exposure. Robins (1999b) accommodated this problem by introducing structural nested direct-effects models. Their direct effect parameters can be estimated using inverse probability weighting by a conditional distribution of the mediator. The resulting estimators are consistent, but inefficient and can be extremely unstable when the intermediate variable is absolutely continuous, because minor errors in the density of the mediator (e.g. due to random noise or model misspecification) may get severely inflated in the inverse weighting procedure. In this chapter, we develop direct effect estimators which are not only more efficient, but also consistent under a less

demanding model for a conditional expectation of the outcome. We find the one estimator which avoids inverse probability weighting altoghether by using sequential G-estimation to perform best. This estimator is intuitive, computationally straightforward and, as demonstrated by simulation, competes extremely well with ordinary least squares estimators in settings where standard regression is valid.

[Original co-author: S. Vansteelandt and E. Goetghebeur]

## 6.1 Introduction

Once researchers have established that an exposure affects an outcome, the attention typically turns to understanding the biologic/mechanistic pathways that contribute to this effect. Empirically, this is most naturally approached by disentangling the part of the exposure effect that is explained by intermediate effects of the exposure on the outcome through given mediators, and by the remaining direct effect. The following examples illustrate this.

(*Example 1: Surrogate biomarkers*) The pressure of accelerated evaluation of new AIDS therapies has led to the use of CD4 blood count and viral load as endpoints that replace time to clinical events and overall survival. This raises the question whether an effect of treatment on the biomarker provides evidence for a clinical effect (Molenberghs et al., 2004). While a good biomarker need not lie on the causal path from treatment to clinical event, a biomarker which does, is often more trustworthy. A number of approaches have therefore been developed to infer whether the effect of treatment on the outcome is entirely mediated by its effect on the biomarker (Frangakis and Rubin, 2002; Taylor et al., 2005). These approaches are of special interest in settings where data from a single study are available and prediction-based approaches (Molenberghs et al., 2004) are thus not applicable.

(*Example 2: Gender discrimination*) Bickel, Hammel and O'Connell (1975) examine data on sex bias in university graduate admissions. Noting

that study choices are on average different between male and female applicants, the investigation of gender discrimination may be approached by evaluating whether there is a direct gender effect on admission rates, which is not mediated by study choice.

(*Example 3: Zygosity in reproductive epidemiology*) Verstraelen et al. (2005) estimate that the odds of preterm birth in twins conceived after in vitro fertilization (IVF) is higher than in naturally conceived twins, after controlling for maternal age and parity. Since many more twins conceived after subfertility treatment are dizygotic and since perinatal outcomes tend to be better for dizygous than for monozygous twins, this effect of IVF on birth weight is partly explained by its effect on zygosity. Verstraelen et al. (2005) thus infer the effect which subfertility treatment has on preterm birth risk, other than through modifying the dizogytic/monozygotic twinning rate.

Traditional regression approaches for direct effects estimate the residual exposure effect that remains on the outcome after adjusting for the given mediator. These approaches tend to be biased by the same token that adjustment for post-randomization measurements may introduce bias in the analysis of randomized experiments (Rosenbaum, 1984). This is so whenever there exist common causes of the mediator and outcome, other than the considered exposure (Cole and Hernan, 2002; Pearl, 2000; Robins, 1986). In some cases, the absence of such common causes may be accepted based on biological grounds. For instance, Verstraelen et al. (2005) used standard adjustment for zygosity to estimate the direct effect of subfertility treatment on preterm birth because it is reasonable to assume that zygosity is not affected by risk factors of preterm birth other than subfertility treatment (and parental fertility) itself. When, as usually, the presence of common causes of mediator and outcome cannot be precluded, as in most cases of interest, untestable assumptions must be made. In this chapter, as in Robins (1999b) and Petersen, Sinisi and van der Laan (2006), we proceed under the assumption of no unmeasured confounders for the association between mediator and outcome. Intuitively, this assumption is sufficient because the size of the direct effect depends on how strongly the mediator affects the outcome and inferring the latter requires knowing all

common causes of both mediator and outcome. Ten Have et al. (2007) avoid this assumption but assume instead that exposure and mediator do not interact in their effect on the outcome, and that the effect of exposure on the mediator varies by baseline covariates. However, this method typically comes with large standard errors for the estimated effects and thus, with information loss.

Even when all confounders for the association between mediator and outcome have been measured, standard regression adjustment is not valid for estimating the direct effect of exposure on outcome. It is prone to bias whenever some of these confounders are themselves affected by the treatment. This happens for the same reason that stratifying by the mediator may induce selection bias. van der Laan and Petersen (2005) and Robins (1999b) accommodate this via inverse probability of treatment weighting estimators for the parameters indexing marginal structural models and structural nested direct effects models, respectively. Both classes of estimators involve inverse probability weighting by a conditional distribution of the mediator. As demonstrated by extensive simulation studies in Section 6.3, these estimators can be extremely inefficient and unstable when there are strong predictors of the mediators, or when the mediator is absolutely continuous; in the latter case, they are also likely biased by the fact that models for a conditional density are difficult to postulate.

In this chapter, we mitigate these problems by developing estimators for the direct effect parameters indexing structural nested direct effects models, which are asymptotically unbiased as soon as a less demanding model for the conditional expectation of the outcome is correctly specified. One of the estimators avoids inverse probability weighting altoghether by using a sequential G-estimation procedure. This estimator is intuitive and computationally straightforward. As demonstrated by extensive simulation studies, it competes extremely well with standard ordinary least squares estimators in settings where standard regression is valid, but in contrast, remains valid when some of the considered confounders are themselves affected by the exposure. Our methods also provide insights on how to stabilize estimators based on inverse probability weighting in the presence of extreme weights.

## 6.2 Structural nested direct effect models

### 6.2.1 Controlled direct effects

Let $Y_{xk}$ be the potential outcome which a given subject would have experienced under exposure $X = x$ and a fixed value $k$ for the intermediate variable $K$. Then, as in Robins (1999b), we formally define the direct effect on outcome $Y$ of setting exposure $X = x$ (versus $X = 0$), when holding $K$ fixed, as the contrast $Y_{xk} - Y_{0k}$ between the two potential outcomes $Y_{xk}$ and $Y_{0k}$ for the same subject. This is termed a controlled direct effect. In this chapter, we develop inference for structural nested direct effect (SNDE) models (Robins, 1999b) which parameterize average controlled direct effects conditionally on pre-exposure covariates $S$ and among subjects with $X = x$:

$$E(Y_{xk} - Y_{0k}|X = x, S) = m(x, k, S; \psi^*) \tag{6.1}$$

where $m(x, k, S; \psi)$ is a known function, smooth in $\psi$, satisfying $m(0, k, S; \psi) = 0$ and where $\psi^*$ is an unknown finite-dimensional parameter. For example, assuming that the direct effect of exposure $x$ (versus 0) is linear in $x$ and the same regardless of $k$ and $S$, we may choose $m(x, k, S; \psi) = \psi x$.

### 6.2.2 Inverse Probability of Intermediate Weighted estimators

Inference for $\psi^*$ in model (6.1) is developed by Robins (1999b) and briefly reviewed here from a different perspective. Suppose first that the potential outcome $Y_k \equiv Y_{Xk}$ following setting $K = k$ is observed for every subject and every value $k$ on the support of $K$. Further, assume that, as expressed by the causal diagram of Figure 6.1, $S$ contains all confounders for the association between $X$ and $Y_k$ so that

$$Y_{xk} \amalg X|S \qquad \forall(x, k) \tag{6.2}$$

Then, for each $k$, model (6.1) is a structural nested mean model (Robins, 1994) which can be fitted by G-estimation (Robins, Mark and Newey, 1992). That is, $\psi^*$ can be estimated as the value $\psi$ such that, after subtracting

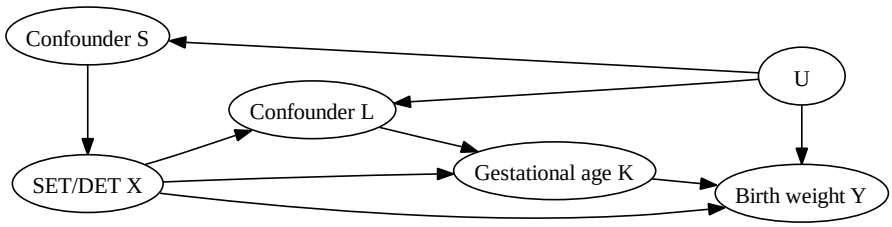


Figure 6.1: *Causal Diagram*

the direct effect $m(X, k, S; \psi)$ from $Y_k$, no dependence on $X$ remains, conditionally on $S$. Specifically, for given $k$, all unbiased estimating functions for $\psi^*$ in the model given by restrictions (6.1) and (6.2) for the given $k$, with $Y_k$ observed, are of the form

$$\Delta\{d_k(X, S)|S\}\{Y_k - m(X, k, S; \psi) - q_k(S)\} \tag{6.3}$$

where $d_k(X, S)$ is an arbitrary vector function of the dimension of $\psi$, $q_k(S)$ is an arbitrary scalar function and where for any 2 random variables $A$ and $B$, we define $\Delta\{A|B\} \equiv A - E(A|B)$. For example, we may choose $d_k(X, S) = X$, which, as we will show later, corresponds to the optimal choice for $d_k(X, S)$ when model (6.1) is linear in $x$ and independent of $k$ and $S$ (i.e. $m(X, k, S; \psi) = \psi X$). That (6.3) is unbiased at $\psi = \psi^*$ follows because $E(Y_k - m(X, k, S; \psi)|X, S) = E(Y_{0k}|X, S) = E(Y_{0k}|S)$ under the model given by restrictions (6.1) and (6.2). It then follows that all unbiased estimating functions for $\psi^*$ in model (6.1)-(6.2) (for all $k$) with $Y_k$ observed are of the form

$$\int \Delta\{d_k(X, S)|S\}\{Y_k - m(X, k, S; \psi) - q_k(S)\}\, dk \tag{6.4}$$

Estimating equations based on (6.4) yield no feasible estimators for $\psi^*$ because $Y_k$ is unknown for each $k$ except the observed realization of $K$. Multiplying each term in (6.4) with $I(K = k)$ yields an observed data estimating function, which in general no longer has mean zero because subjects with $K = k$ may form a selective subgroup. To correct for this, we make the additional assumption, which is expressed by the diagram

of Figure 6.1, that $(X, L, S)$ contains all confounders for the association between $K$ and $Y$ so that

$$Y_{xk} \amalg K | X = x, L, S \qquad \forall x, k \tag{6.5}$$

This assumption allows for inversely weighting each term in the estimating function (6.4) by the conditional distribution $f(K|L, S, X)$ of $K$ given $X$, $L$ and $S$, as in

$$\begin{aligned} &\int \frac{I(K=k)}{f(K=k|L,S,X)} \Delta\{d_k(X,S)|S\} \{Y_k - m(X,k,S;\psi) - q_k(S)\} \, dk \\ &= \frac{\Delta\{d_K(X,S)|S\}}{f(K|L,S,X)} \{Y - m(X,K,S;\psi) - q_K(S)\} \end{aligned} \tag{6.6}$$

Estimating function (6.6) has mean zero at $\psi = \psi^*$ under the model defined by (6.1), (6.2) and (6.5) because the conditional mean of

$$\frac{I(K=k)}{f(K=k|L,S,X)}$$

given $(L, S, X, Y_k)$ equals 1. This unbiasedness is key to the fact that the solution to the estimating equation

$$0 = \sum_{i=1}^{n} \frac{\Delta\{d_{K_i}(X_i,S_i)|S_i\}}{f(K_i|L_i,S_i,X_i)} \{Y_i - m(X_i,K_i,S_i;\psi) - q_{K_i}(S_i)\} \tag{6.7}$$

is (under standard, weak regularity conditions) a consistent and asymptotically normal (CAN) estimator of $\psi^*$, provided that $f(K = k|L, S, X) > 0$ with probability 1 for all $k$ in the support of $K$ (Robins, 1999b).

Solving (6.7) requires that we specify parametric models

$$f(K|L,S,X) = f(K|L,S,X;\alpha^*) \tag{6.8}$$

$$E(d_K(X,S)|S) = \int d_K(X,S) f(X|S) dX = E(d_K(X,S)|S;\beta^*) \tag{6.9}$$

where $f(K|L, S, X; \alpha)$ is a conditional density function, smooth in $\alpha$, $E(d_K(X, S)|S; \beta)$ is a function of $S$, smooth in $\beta$, and $(\alpha^*, \beta^*)$ is an unknown finite-dimensional parameter. For example, we may assume that

the conditional distribution of $K$ given $(L, S, X)$ is normal with mean $\alpha_0 + \alpha_1 L + \alpha_2 S + \alpha_3 X$ and constant standard deviation $\sigma_K$ and, with $d_K(X, S) = X$, that $E(X|S; \beta) = \beta_0 + \beta_1 S$.

Throughout we let $\mathcal{A}$ be the model for the observed data defined by the model restrictions (6.1), (6.8) and (6.9), and the no unmeasured confounders assumptions (6.2) and (6.5). Let $\hat{\alpha}$ and $\hat{\beta}$ be (root-$n$) consistent estimators for $\alpha^*$ and $\beta^*$, respectively, such as can be obtained via standard regression. It then follows from the previous discussion that a CAN estimator $\hat{\psi}_{IPIW}$ for the direct effect parameter $\psi^*$ under model $\mathcal{A}$ can be obtained by solving

$$0 = \sum_{i=1}^{n} U_{i,IPIW}(d, q; \psi, \hat{\alpha}, \hat{\beta}) \tag{6.10}$$

where

$$U_{i,IPIW}(d, q; \psi, \alpha, \beta) = \frac{\Delta\{d_{K_i}(X_i, S_i)|S_i; \beta\}}{f(K_i|L_i, S_i, X_i; \alpha)} \{Y_i - m(X_i, K_i, S_i; \psi) - q_{K_i}(S_i)\} \tag{6.11}$$

For given $k$, optimal choices for $d_k(X, S)$ and $q_k(S)$ which lead to a semi-parametric efficient estimator of $\psi^*$ in the model given by restrictions (6.1) and (6.2) (for the given $k$) and with $Y_k$ observed, have been derived by Robins (1994). When the potential outcome variance $\mathrm{Var}(Y_k|X, S)$ is constant in $(X, S)$, these choices equal

$$\begin{aligned} d_k(X, S) &= \frac{\partial m(X, k, S; \psi)}{\partial \psi} \\ q_k(S) &= E\{Y_k - m(X, k, S; \psi)|S\} \end{aligned} \tag{6.12}$$

where the latter can be calculated using the law of iterated expectations

$$E(Y_k - m(X, k, S; \psi)|S) = E\left[E(Y|K = k, X, L, S) - m(X, k, S; \psi)|S)\right]$$

The same choices may not lead to a semi-parametric efficient estimator of $\psi^*$ under model $\mathcal{A}$, in which $Y_k$ is not observed for each subject. However, we recommend using the above choices for practical use because we

conjecture that they will generally yield reasonable efficiency, while calculating the semi-parametric efficient estimator under model $\mathcal{A}$ is much more tedious as it requires solving integral equations. In addition, using the choice (6.12) yields the additional advantage that the estimator $\hat{\psi}_{IPIW}$, and likewise all other estimators developed in this article, remains CAN when instead of model (6.9), a model for the conditional expectation (6.12) is correctly specified (for each $k$). Our focus on correct specification of (6.9) is motivated by the fact that this model is usually known exactly when $X$ is an exposure that is randomly assigned, conditional on $S$.

When the intermediate variable is absolutely continuous, the above method requires inverse weighting by a density. The inverse weighting estimator $\hat{\psi}_{IPIW}$ is then likely to have serious finite sample bias because statistical models for a density are difficult to postulate and small misspecifications in the tails of the density can have a large effect on the direct-effects estimates through their influence on the inverse weights. Furthermore, the large variability of the inverse weights may then seriously distort the precision of the estimate. An ad hoc approach to stabilize the inverse weights is to multiply the estimating function (6.6) by $f(K|S)$, because observations with extreme values for $f(K|L,S,X)$ are likely also extreme in terms of $f(K|S)$ and may therefore have a more stable ratio of both. The resulting estimating function remains unbiased because $d_K(X,S)$ is an arbitrary function of $K, X$ and $S$, and the conditional expectation in $E(d_K(X,S)|S)$ is only w.r.t. $X$. Therefore, from now on, we will replace the weights $1/f(K|L,S,X)$ by the stabilized weights $f(K|S)/f(K|L,S,X)$. However, as we will show in several simulation studies in Section 6.3, this ad hoc stabilization will often not suffice to obtain well-behaved estimators in moderate sample sizes. Alternatively, one could truncate the weights (Wang et al., 2006). However, one may argue that truncated weights are deliberately misspecified weights and, as such, may impact the consistency of the direct-effects estimator. In the next sections, we will therefore develop estimators which allow misspecification (and thus truncation) of the weights.

### 6.2.3 Doubly-robust estimators

To obtain estimators with better performance in the presence of unstable weights, note (using similar arguments as in van der Laan and Robins, 2003) that, up to asymptotic equivalence, all CAN estimators for $\psi^*$ under model $\mathcal{A}$ can be obtained by solving estimating equations of the form

$$0 = \sum_{i=1}^{n} U_{i,IPIW}(d, q; \psi, \alpha, \beta) - \Delta\left\{\phi(K_i, L_i, X_i, S_i)|L_i, X_i, S_i\right\} \quad (6.13)$$

where $\phi(K_i, L_i, X_i, S_i)$ is an arbitrary vector function of the dimension of $\psi$. Part 2 of Theorem 5 below shows that for given $d_K(X, S)$ and $q_K(S)$, the optimal choice of $\phi(K_i, L_i, X_i, S_i)$ that leads to estimators of $\psi^*$ with minimum asymptotic variance, equals

$$\phi_{opt}(K_i, L_i, X_i, S_i) \equiv E(U_{i,IPIW}(d, q; \psi, \alpha, \beta)|K_i, L_i, X_i, S_i) \quad (6.14)$$

In the proof of Theorem 5 (see Appendix 6.S1), we further show that this yields the following estimating function for $\psi$

$$\Delta\{d_K(X, S)|S; \beta\}W(\alpha)\Delta\left\{Y|K, L, X, S\right\} +$$

$$\int \Delta\{d_K(X, S)|S; \beta\}\left\{E(Y|K, L, X, S) - m(X, K, S; \psi) - q_K(S)\right\} f(K|S)dK \quad (6.15)$$

where $W(\alpha) = f(K|S)/f(K|L, S, X; \alpha)$ and where the conditional density $f(K|S)$ may be replaced by an estimate. Using this estimating function requires that we specify a parametric model

$$E(Y|K, L, X, S) \quad = \quad E(Y|K, L, X, S; \gamma^*) \quad (6.16)$$

where $E(Y|K, L, X, S; \gamma)$ is a function of $(K, L, X, S)$, smooth in $\gamma$, and $\gamma^*$ is an unknown finite-dimensional parameter. A consistent estimator $\hat{\gamma}$ for $\gamma^*$ can be obtained using standard regression techniques. For example, for the linear model

$$E(Y|K, L, X, S; \gamma) = \gamma_0 + \gamma_1 K + \gamma_2 L + \gamma_3 X + \gamma_3 S \quad (6.17)$$

and with $m(X, S, K; \psi) = \psi X$, $d_K(X, S) = X$ and using the optimal choice of $q_K(S)$, the estimating function (6.15) has the relatively simple form

$$\Delta\{X|S;\beta\}\left[W(\alpha)\Delta\{Y|K,L,X,S;\gamma\} + \gamma_2\Delta\{L|S\} + (\gamma_3 - \psi)\Delta\{X|S\}\right]$$

Part 1 of Theorem 5 shows that the solution $\hat{\psi}_{DR}$ to an estimating equation based on (6.15) has the interesting feature of being a consistent estimator of $\psi^*$ when either model (6.8) holds or model (6.16), but not necessarily both. We therefore call $\hat{\psi}_{DR}$ a doubly-robust estimator of $\psi^*$.

**Theorem 5.** *1. The solution $\hat{\psi}_{DR}$ to equation*

$$0 = \sum_{i=1}^{n} U_{i,DR}(d, q; \psi, \hat{\alpha}, \hat{\beta}, \hat{\gamma}) \tag{6.18}$$

*where*

$$\begin{aligned} U_{i,DR}(d, q; \psi, \alpha, \beta, \gamma) \quad = \\ & \Delta\{d_{K_i}(X_i, S_i)|S_i;\beta\}W_i(\alpha)\Delta\left\{Y_i|K_i, L_i, X_i, S_i;\gamma\right\} \\ & + \int \Delta\{d_{K_i}(X_i, S_i)|S_i;\beta\}\left\{E(Y_i|K_i, L_i, X_i, S_i;\gamma)\right. \\ & \left. -m(X_i, K_i, S_i;\psi) - q_{K_i}(S_i)\right\} f(K_i|S_i)dK_i \end{aligned} \tag{6.19}$$

*is a consistent estimator of $\psi^*$ under model $\mathcal{A} \cup \mathcal{B}$, where $\mathcal{B}$ is the model for the observed data defined by the model restrictions (6.1), (6.9) and (6.16), and the no unmeasured confounders assumptions (6.2) and (6.5).*

*2. Let $\hat{\psi}(\phi)$ be the solution to (6.13) for the given choice of $\phi(K_i, L_i, X_i, S_i)$, for the same choice of $d_K(X, S)$ and $q_K(S)$ as used to obtain $\hat{\psi}_{DR}$, and with $(\alpha, \beta, \gamma)$ replaced by $(\hat{\alpha}, \hat{\beta}, \hat{\gamma})$. When the true distribution of the data lies in the intersection model $\mathcal{A} \cap \mathcal{B}$, then the difference in asymptotic variance of $\hat{\psi}(\phi)$ and $\hat{\psi}_{DR}$ is non-negative.*

### 6.2.4 Unweighted estimators and sequential G-estimators

The attractiveness of the doubly-robust estimator $\hat{\psi}_{DR}$ lies not only in it (typically) being more efficient than the simpler inverse weighting estimator $\hat{\psi}_{IPIW}$. Its main attraction lies in the fact that it avoids reliance on a difficult-to-postulate model for the density of the mediator. Instead, it relies on a model for the expected outcome, which is typically easier to specify. In this section, we completely avoid reliance on the model for the density of the mediator by setting $f(K|L,S,X)$ equal to $f(K|S)$ in the estimating function (6.18) of the doubly-robust estimator. The implication of this is to set all weights equal to 1, which leads to an unweighted estimating equation. The corresponding estimators $\hat{\psi}_{UW}$ solve

$$0 = \sum_{i=1}^{n} U_{i,UW}(d,q;\psi,\hat{\beta},\hat{\gamma}) \tag{6.20}$$

where $U_{i,UW}(d,q;\psi,\beta,\gamma)$ is defined as $U_{i,DR}(d,q;\psi,\alpha,\beta,\gamma)$, but with $W_i(\alpha)$ replaced by 1. For example, when choosing a linear conditional mean model for $Y$ as in (6.17), $m(X,K,S;\psi) = \psi X$, $d_K(X,S) = X$, $q_K(S) = \gamma_1 E(K|S)$, we obtain the simple form

$$0 = \sum_{i=1}^{n} \Delta\{X_i|S_i;\beta\}\,(Y_i - \gamma_1 K_i - \psi X_i) \tag{6.21}$$

Note that this estimating equation is very intuitive as it expresses that, after subtracting the effect $\gamma_1 K_i$ of the mediator and the direct effect $\psi X_i$ of the exposure from the outcome, no association with $X_i$ should remain after adjustment for the confounder $S_i$. As such, the solution $\hat{\psi}$ for $\psi$ to equation (6.21) with $\gamma_1$ replaced by a consistent estimate $\hat{\gamma}_1$, can be viewed as a sequential G-estimator (i.e., it is obtained by G-estimation applied to the residual outcome $Y_i - \hat{\gamma}_1 K_i$ that remains after removing the effect of the mediator from the outcome). The 'unweighted' estimators that solve (6.20) generalize such sequential G-estimators by allowing for nonlinear models (6.16) for the outcome and for SNDE models that incorporate interactions between $X$ and $K$, and by enabling greater efficiency.

By the fact that the solutions to (6.18) are consistent estimators for $\psi^*$ under model $\mathcal{A} \cup \mathcal{B}$, solving (6.20) gives a consistent estimator for $\psi^*$ under

model $\mathcal{B}$. In the simulation study of Section 6.3, we will show that the resulting estimator has the desirable property of being very stable and efficient as a result of avoiding the inverse weighting, but is no longer doubly-robust. In the following sections, we briefly introduce alternative estimators which are designed to perform well in the presence of extreme weights and protect the double robustness property.

### 6.2.5 Stabilized doubly-robust estimators

Using arguments similar to Robins et al. (2007), we will stabilize the doubly-robust direct effects estimator by substituting $\psi$ in expression (6.14) by an estimator $\tilde{\psi}$ which is consistent under model $\mathcal{B}$. We denote the resulting estimator with $\hat{\psi}_{SDR}$. When considering closed-form estimators for $\psi^*$ obtained from expression (6.18), it can be seen that the impact of this is that the weights $W_i(\alpha)$ appear both in the numerator and denominator. For example, with $m(X, K, S; \psi) = \psi X$, $d_K(X, S) = X$, $q_K(S) = 0$ and a linear conditional mean model for $Y$ as in (6.17), we then obtain

$$\begin{aligned}\hat{\psi}_{SDR} &= \frac{\sum_{i=1}^{n} \Delta\{X_i|S_i;\beta\}\left[W_i(\alpha)\left\{\Delta\{Y_i|K_i,L_i,X_i,S_i;\gamma\}+\tilde{\psi}X_i\right\}\right]}{\sum_{i=1}^{n} W_i(\alpha)X_i\Delta\{X_i|S_i;\beta\}} \\ &\quad + \frac{\sum_{i=1}^{n}\left\{\tilde{\psi}X_i - E(Y_i|K_i = E(K_i|S_i), L_i, S_i, X_i;\gamma)\right\}}{\sum_{i=1}^{n} W_i(\alpha)X_i\Delta\{X_i|S_i;\beta\}} \qquad (6.22)\end{aligned}$$

The resulting estimator is generally more stable than the doubly-robust estimator

$$\frac{\sum_{i=1}^{n} \Delta\{X_i|S_i;\beta\}\left[W_i(\alpha)\Delta\{Y_i|K_i,L_i,X_i,S_i;\gamma\}+E(Y_i|K_i=E(K_i|S_i),L_i,S_i,X_i;\gamma)\right]}{\sum_{i=1}^{n} X_i\Delta\{X_i|S_i;\beta\}}$$

of Section 6.2.3 by the fact that subjects with extreme weights $W_i(\alpha)$ in the numerator of (6.22) will also make the denominator of (6.22) extreme. The stabilized doubly-robust estimator $\hat{\psi}_{SDR}$ is a consistent estimator of $\psi^*$ under model $\mathcal{A}$, even when model (6.16) is misspecified and thus even when $\tilde{\psi}$ is an inconsistent estimator, because estimating equation (6.13) is unbiased under model $\mathcal{A}$ regardless of $\phi(K_i, L_i, X_i, S_i)$ (and thus in particular when the unknown parameters indexing $\phi(K_i, L_i, X_i, S_i)$ are replaced

by inconsistent estimators). Likewise, $\hat{\psi}_{SDR}$ is a consistent estimator of $\psi^*$ under model $\mathcal{B}$, even when model (6.8) is misspecified, because estimating equation (6.18) is unbiased under model $\mathcal{B}$ and because $\tilde{\psi}$ is a consistent estimator of $\psi^*$ under model $\mathcal{B}$. It follows that $\hat{\psi}_{SDR}$ is a doubly-robust estimator of $\psi^*$.

Alternatively, we may improve the finite-sample behavior of $\hat{\psi}_{DR}$ by adapting ideas in Tan (2006) for inverse weighting estimators to inverse weighting estimating functions. Specifically, we modify the doubly-robust estimating equation for $\psi^*$ as

$$0 = \sum_{i=1}^{n} U_{i,IPIW}(d, q; \psi, \alpha, \beta) - \kappa \Delta \left\{ \phi_{\text{opt}}(K_i, L_i, X_i, S_i) | L_i, X_i, S_i \right\} \quad (6.23)$$

and determine an 'optimal' choice of $\kappa$ that leads to improved efficiency. Note that the choice $\kappa = 1$ yields the estimator $\hat{\psi}_{DR}$, which may be an inefficient doubly-robust estimator whenever model (6.16) is incorrectly specified.

Let for notational convenience $\xi \equiv \Delta \left\{ \phi_{opt}(K_i, L_i, X_i, S_i) | L_i, X_i, S_i \right\}$ and $\eta \equiv U_{i,IPIW}(d, q; \psi, \alpha, \beta)$. For arbitrary random variable $A$, define $\hat{E}(A)$ as the sample average $\sum_{i=1}^{n} A_i / n$. Then choosing $\kappa$ equal to $\kappa_{opt} = \hat{E}^{-1}(\xi\xi')\hat{E}(\xi\eta')$ yields an estimator $\hat{\psi}(\kappa_{opt})$ with minimal variance among all estimators $\hat{\psi}(\kappa)$ that solve (6.23) for given $\kappa$. This can be seen from the following 2 arguments. First, the variance $E(\eta^2 - 2\kappa\eta\xi + \kappa^2\xi^2)$ of the estimating function $\eta - \kappa\xi$ is minimized at $\kappa_{opt}$. Second, the estimator obtained by solving the corresponding estimating equation itself has minimal variance among all estimators $\hat{\psi}(\kappa)$ because

$$Var(\hat{\psi}(\kappa)) \approx \frac{1}{n} E\left(\frac{\partial \eta}{\partial \psi}\right)^{-1} Var\left(\eta - \kappa\xi\right) E\left(\frac{\partial \eta}{\partial \psi}\right)^{-1'}$$

and thus the variance of these estimators is proportional to the variance of their estimating function. The estimator $\hat{\psi}(\kappa_{opt})$ is however not doubly-robust because $\kappa_{opt}$ may not converge to 1 under a correctly specified model for (6.16).

Choosing $\kappa$ to equal

$$\kappa_{\text{dr}} \equiv \hat{E}^{-1}(\xi\chi')\hat{E}(\xi\eta')$$

with

$$\chi \equiv \Delta\{d_{K_i}(X_i, S_i)|S_i; \beta\}W_i(\alpha)\left[E\left\{Y_i - m(X_i, K_i, S_i; \psi)|X_i, K_i, L_i, S_i\right\} - q_{K_i}(S_i)\right]$$

accommodates this. Indeed, $\kappa_{\rm dr}$ converges to 1 when the model for (6.16) is correctly specified, because $\chi = E(\eta|X, K, L, S)$ and thus $E^{-1}(\xi\chi')E(\xi\eta') = 1$ under such correctly specified model. It follows that the estimator $\hat{\psi}(\kappa_{\rm dr})$ is a doubly-robust estimator. Further,

$$E(\xi\chi') \;\; = \;\; E\left[\xi\left\{\xi + E(\eta|X, L, S)\right\}'\right] = E(\xi\xi')$$

if model (6.8) is correctly specified, because $\xi$ has conditional mean zero, given $(X, L, S)$, under that model. It follows that $\kappa_{\rm dr} - \kappa_{opt}$ converges to zero under model (6.8), suggesting that $\hat{\psi}(\kappa_{\rm dr})$ has minimal variance among all estimators $\hat{\psi}(\kappa)$ under that model. Throughout this chapter, we will refer to $\hat{\psi}(\kappa_{\rm dr}) \equiv \hat{\psi}_{IDR}$ as an improved doubly-robust estimator.

Finally, combining the ideas leading to the estimators $\hat{\psi}_{SDR}$ and $\hat{\psi}_{IDR}$ leads to yet a final estimator that we will refer to as the stabilized, improved doubly-robust estimator. The resulting estimator is obtained by substituting $\kappa$ with $\kappa_{\rm dr}$ in (6.23) and $\psi$ in expression (6.14) by an estimator $\tilde{\psi}$ which is consistent under model $\mathcal{B}$. We will denote it as $\hat{\psi}_{SIDR}$.

## 6.3 Simulation study

We generate 1000 datasets of size 1500 according to the data generating mechanism of Figure 6.1, but without confounder $S$. All analyses were conducted in R (version 2.3.1). In a first simulation experiment, we postulate linear models for all variables in the diagram: $X = 1 + \epsilon_X$, $L = 1+\lambda X+0.8U+\epsilon_L$, $K = 0.5L-0.5X+\epsilon_K$ and $Y = \delta(-1+2X+0.5K+U+\epsilon_Y)$ for mutually independent, normally distributed variates $U$, $\epsilon_X$, $\epsilon_L$, $\epsilon_K$ and $\epsilon_Y$ with mean zero and standard deviations 1, 0.5, 1, 0.3 and 0.5, respectively and with $\delta = 1$. We considered both the cases $\lambda = 1.5$ and $\lambda = 0$ to represent settings where $L$ is/is not affected by $X$. As such, we represent both settings where standard regression methods are/are not applicable for estimating the direct effect (i.e. $2\delta$) of $X$ on $Y$ (which is not mediated by $K$). A characteristic feature of the simulation experiments with

$\lambda = 1.5$ is that there is a strong association between $X$ and $Y$ along the path $X - L - U - Y$.

Assuming a correctly specified structural nested direct-effects model with $m(X,$
$K; \psi) = \psi X$, the following estimators were calculated in each simulation, corresponding to the choices $d_K(X) = X$ and $q_K = 0$: the Inverse Probability of Intermediate Weighting (IPIW) estimator of Section 6.2.2, the doubly-robust (DR) estimator of Section 6.2.3, the sequential G-estimator (SG) of Section 6.2.4, the stabilized doubly-robust (SDR) estimator, the improved doubly-robust (IDR) estimator and the stabilized, improved doubly-robust (SIDR) estimator of Section 6.2.5. The sequential G-estimator was used as a preliminary estimator ($\tilde{\psi}$) in both stabilized estimators. Finally, we also reported the estimated coefficient for $X$ in a linear regression model for $Y$, given $X, K$ and $L$. We chose the following correctly specified working models: a normal conditional distribution for $K$ given $L$ and $X$ with mean linear in $L$ and $X$ and constant residual standard deviation, and a linear regression model for $Y$ with mean linear in $X$, $K$ and $L$.

Because of outlying values for a number of estimators, Tables 6.1-6.3 reports both the average and median bias, the average and median bootstrap standard error, the empirical standard deviation of the estimates and corresponding (robust) Minimum Covariance Determinant (MCD) estimator for the standard deviation, the $p$-value of the Wilcoxon rank test whether the median direct-effect estimate differs from zero, and the coverage of standard 95% bootstrap confidence intervals. Here, bootstrap estimates are based on 1000 bootstrap samples.

In the first simulation experiment (see Table 6.1), we find that the standard linear regression analysis (LM) yields severely biased estimates when the confounder $L$ is affected by $X$, while all other estimators are approximately unbiased. The IPIW estimator is unstable in the sense that it suffers from many outlying values. The DR estimator is considerably more stable and more efficient. Slightly higher efficiency is observed for the improved doubly-robust estimator, but the best results are obtained using the unweighted estimator. The simulation experiment where $L$ is not affected by $X$ reveal that the latter estimator competes very well with the

standard regression analysis. Indeed, it is only slightly less efficient, but has the advantage of remaining unbiased when $L$ is affected by $X$.

In a second simulation experiment (see Table 6.2), we investigate the impact of model misspecification by generating $K$ as $\exp(0.5L-0.5X+\epsilon_K)$, with all remaining variables generated as before. Estimators were obtained using the same working models that were previously used. We now obtain extremely unstable IPIW and DR estimates as a result of the estimated density of the intermediate taking extremely small values for some subjects (due to the skewness of the data). The improved doubly-robust estimators perform considerably better, with the stabilized improved DR estimator having the best performance. However, as a result of remaining instability, bootstrap standard errors could not be obtained in all simulated datasets. Overall, the most efficient estimates are again obtained via the sequential G-estimator.

In a third simulation experiment (see Table 6.3), we misspecified working model (6.16) by generating $Y$ as $Y = \delta(-1+2X+0.5(K-E(K))-3(K-E(K))^2 + U + \epsilon_Y)$, with $\delta = 0.7$ to obtain the same variability in $Y$ as in the first simulation expiriment. Results are now similar to those of the first simulation experiment. Curiously, also the sequential G-estimator, while fully relying on the misspecified working model (6.16), remains unbiased. When $L$ is not affected by $X$, this can be understood from the following arguments. Fitting the outcome model (6.17) (with $S$ empty) then yields valid estimates for the direct effect $\psi^*$ of $X$ on $Y$, even when the association between $K$ and $Y$ is misspecified, because the conditional mean of $X$ is linear in $K$ and $L$ under the considered data-generating mechanism (see e.g. Robins, Mark and Newey, 1992). From the form of the normal equations for the parameters indexing model (6.17), it thus follows that $\Delta\{X\}(Y - \gamma_0^* - \gamma_1^* K - \psi^* X - \gamma_2^* L)$, with $\gamma_0^*, \gamma_1^*$ and $\gamma_3^*$ the limiting values of the ordinary least squares estimators for $\gamma_0, \gamma_1$ and $\gamma_3$ under model (6.17), has mean zero. In particular, because $L$ is not affected by $X$ and thus independent of $X$ under our model, we have that the estimating function of the sequential G-estimator,

$$\Delta(X)(Y - \gamma_1^* K - \psi^* X),$$

| with effect of $X$ on $L$ | | | | | |
|---|---|---|---|---|---|
| | bias (mean/med) | boot SE (mean/med) | SE (emp/MCD) | p-value bias= 0 | boot cov |
| IPIW | -0.49/-0.024 | 44.98/0.20 | 13.59/0.28 | 0.01 | 92.3 |
| DR | 0.029/-0.001 | 1.36/0.12 | 1.37/0.17 | 0.73 | 96.1 |
| UW | -0.001/-0.001 | 0.061/0.061 | 0.061/0.061 | 0.51 | 95.0 |
| SDR | -0.055/0.00 | 26.53/0.15 | 1.91/0.21 | 0.51 | 94.1 |
| IDR | -0.011/-0.010 | 0.16/0.10 | 0.21/0.13 | 0.11 | 93.9 |
| SIDR | -0.016/-0.006 | 1.013/0.12 | 0.40/0.15 | 0.43 | 95.3 |
| LM | -0.73/-0.73 | 0.068/0.068 | 0.071/0.072 | 0.00 | 0 |
| without effect of $X$ on $L$ | | | | | |
| | bias (mean/med) | boot SE (mean/med) | SE (emp/MCD) | p-value bias= 0 | boot cov |
| IPIW | 0.18/0.032 | 49.33/0.26 | 1.79/0.37 | 0.00 | 94.4 |
| DR | 0.009/0.005 | 1.23/0.12 | 1.25/0.18 | 0.60 | 95.5 |
| UW | 0.004/0.007 | 0.07/0.07 | 0.071/0.071 | 0.059 | 95.3 |
| SDR | 0.034/0.002 | 18.92/0.16 | 2.59/0.23 | 0.71 | 94.0 |
| IDR | 0.005/0.002 | 0.15/0.10 | 0.17/0.13 | 0.19 | 94.7 |
| SIDR | -0.003/0.006 | 1.88/0.13 | 0.24/0.16 | 0.28 | 95.3 |
| LM | 0.003/0.003 | 0.062/0.062 | 0.063/0.064 | 0.078 | 95.3 |

Table 6.1: *Results of Simulation Experiment 1*

is unbiased, even when the association between $K$ and $Y$ is misspecified. It can be seen with some algebra that this result continues to hold when $L$ is affected by $X$ and $(X, K, L)$ is multivariate normal.

In a fourth simulation experiment (see Table 6.4), we misspecified both working models by generating $K$ and $Y$ as in the previous 2 simulation experiments, respectively, but with $\delta = 0.04$. As expected, all estimators are now biased. Note that, while the additional misspecification of the working model (6.8) has no immediate impact on the sequential G-estimator (because this estimator avoids inverse probability weighting), it also becomes biased because the robustness property of this estimator (see previous paragraph) only holds for linear models. Note however, that the sequential G-estimator is still outperforming the other estimators both in terms of precision and bias. With a correctly specified working model (6.8) for the intermediate, but a misspecified outcome model (see Table 6.5, simulation experiment 5), as expected, the sequential G-estimator continues to behave poorly as it does not make use of the working model for the intermediate. However, the (stabilized) improved doubly-robust estimators now outperform the others, both in terms of bias and precision (but the bootstrap confidence intervals are poor in terms of coverage). The usefulness of the latter estimators is most apparent when the intermediate is non-normal and, additionally, this is acknowledged via the working model (6.8). We conjecture that these stabilized doubly-robust estimators will be more competitive with the sequential G-estimator in settings where the weights are more stable, such as may happen when the mediator is binary.

| with effect of $X$ on $L$ | | | | | |
|---|---|---|---|---|---|
| | bias (mean/med) | boot SE (mean/med) | SE (emp/MCD) | p-value bias= 0 | boot cov |
| IPIW | 10.82/9.1 | 437/10.0 | 23.8/5.3 | 0.00 | 60.0 |
| DR | 8.3 $10^{50}$/-0.13 | 1.4 $10^{64}$/1.8 $10^{9}$ | 2.7 $10^{51}$/1.9 | 0.59 | 100.0 |
| UW | -0.001/-0.001 | 0.059/0.059 | 0.059/0.059 | 0.78 | 95.1 |
| SDR | 0.00/0.018 | 41.3/0.97 | 2.1/0.77 | 0.80 | 94.2 |
| IDR | 17.1/0.037 | -/- | 520/1.8 | 0.00 | - |
| SIDR | -0.022/-0.002 | -/- | 0.58/0.088 | 0.86 | - |
| LM | -0.73/-0.73 | 0.061/0.061 | 0.063/0.063 | 0.00 | 0 |
| without effect of $X$ on $L$ | | | | | |
| | bias (mean/med) | boot SE (mean/med) | SE (emp/MCD) | p-value bias= 0 | boot cov |
| IPIW | -15.0/10.2 | 4153/39.0 | 936.4/16.2 | 0 | 84.8 |
| DR | $-3.3/10^{50}$/-4.4 | 2.5 $10^{65}$/8.9 $10^{11}$ | 1.1 $10^{53}$/- | 0.012 | 100.0 |
| UW | 0.002/0.000 | 0.062/0.062 | 0.062/0.062 | 0.23 | 96 |
| SDR | -6.7/0.022 | 594/4.9 | 151.8/2.4 | 0.88 | 96.6 |
| IDR | 3.6/0.033 | -/- | 353.62/2.60 | 0.001 | - |
| SIDR | -0.074/0.005 | -/- | 3.13/0.12 | 0.17 | - |
| LM | 0.002/0.001 | 0.053/0.053 | 0.053/0.053 | 0.41 | 95.3 |

Table 6.2: *Results of Simulation Experiment 2*

| with effect of $X$ on $L$ | | | | | |
|---|---|---|---|---|---|
| | bias (mean/med) | boot SE (mean/med) | SE (emp/MCD) | p-value bias= 0 | boot cov |
| IPIW | -0.63/-0.027 | 84.1/0.33 | 20.8/0.43 | 0.9 | 91.7 |
| DR | -0.017/-0.010 | 1.9/0.17 | 2.0/0.25 | 0.94 | 96.1 |
| UW | -0.002/-0.004 | 0.096/0.095 | 0.099/0.098 | 0.35 | 93.0 |
| SDR | 0.11/-0.008 | 44.6/0.21 | 3.7/0.31 | 0.55 | 94.7 |
| IDR | -0.004/-0.007 | 0.23/0.13 | 0.26/0.16 | 0.45 | 95.1 |
| SIDR | 0.059/-0.008 | 0.53/0.15 | 1.8/0.20 | 0.95 | 95.2 |
| LM | -0.51/-0.52 | 0.12/0.12 | 0.12/0.12 | 0.00 | 1.2 |
| without effect of $X$ on $L$ | | | | | |
| | bias (mean/med) | boot SE (mean/med) | SE (emp/MCD) | p-value bias= 0 | boot cov |
| IPIW | 0.068/-0.004 | 106/0.35 | 7.5/0.51 | 0.01 | 92.9 |
| DR | -0.035/-0.005 | 2.4/0.18 | 2.4/0.26 | 0.66 | 95.1 |
| UW | 0.001/0.003 | 0.12/0.12 | 0.12/0.12 | 0.63 | 95.6 |
| SDR | -0.17/-0.014 | 56.9/0.23 | 3.02/0.35 | 0.22 | 94.5 |
| IDR | -0.008/0.001 | 0.23/0.14 | 0.29/0.18 | 0.83 | 94.6 |
| SIDR | -0.008/-0.007 | 0.65/0.17 | 0.50/0.22 | 0.97 | 95.3 |
| LM | 0.000/0.002 | 0.12/0.12 | 0.12/0.12 | 0.80 | 95.5 |

Table 6.3: *Results of Simulation Experiment 3*

| with effect of $X$ on $L$ | | | | | |
|---|---|---|---|---|---|
| | bias (mean/med) | boot SE (mean/med) | SE (emp/MCD) | p-value bias= 0 | boot cov |
| IPIW | -92.87/-37.12 | 8590.71/51.34 | 405.46/63.32 | 0.00 | 53.2 |
| DR | $-8.210^{52}$ / $-42581$ | $1.3\ 10^{66}$ / $3.5\ 10^{43}$ | $2.0\ 10^{54}$/- | 0.0 | 100.0 |
| UW | 1.85/0.58 | 0.067/0.066 | 0.067/0.066 | 0.00 | 0.0 |
| SDR | -1.65/-20.01 | 493.06/2.70 | 40.03/1.96 | 0.00 | 95.7 |
| IDR | -16.75/0.52 | 224.65/113.73 | 319.29/0.11 | 0.57 | 96.7 |
| SIDR | 1.83/0.55 | 3.49/0.14 | 0.44/0.083 | 0.00 | 93.1 |
| LM | 1.78/-1.16 | 0.054/0.055 | 0.065/0.064 | 0.00 | 0 |
| without effect of $X$ on $L$ | | | | | |
| | bias (mean/med) | boot SE (mean/med) | SE (emp/MCD) | p-value bias= 0 | boot cov |
| IPIW | 16.6/-19.9 | 7999/65.6 | 1507/39.5 | 0 | 79.6 |
| DR | $4.2\ 10^{52}$/2211408 | $3.0\ 10^{67}$/ $1.2\ 10^{13}$ | $1.2\ 10^{51}$ / 2.37 | 0.00 | 100.0 |
| UW | -0.34/-0.33 | 0.11/0.097 | 0.15/0.095 | 0.00 | 0 |
| SDR | 9.4/-11.2 | 3514/0.097 | 826/21.9 | 0.00 | 76.2 |
| IDR | 6.6/0.52 | -/- | 107/1.78 | 0.00 | - |
| SIDR | -0.24/0.55 | -/- | 1.3/0.12 | 0.00 | - |
| LM | -0.34/-0.30 | 0.16/0.14 | 0.15/0.089 | 0.00 | 27.3 |

Table 6.4: *Results of Simulation Experiment 4*

| with effect of $X$ on $L$ | | | | | |
|---|---|---|---|---|---|
| | bias (mean/med) | boot SE (mean/med) | SE (emp/MCD) | p-value bias= 0 | boot cov |
| IPIW | -0.30/-0.035 | 13.68/0.091 | 5.79/0.092 | 0.00 | 92.7 |
| DR | 0.43/0.46 | 1.13/0.35 | 5.04/0.36 | 0.00 | 61.5 |
| UW | 0.64/0.57 | 0.18/0.14 | 0.29/0.15 | 0.00 | 0 |
| SDR | 201/0.41 | 137.52/0.43 | 49.38/0.47 | 0.00 | 70.3 |
| IDR | 0.042/0.0014 | 0.13/0.097 | 0.22/0.11 | 0.57 | 74.7 |
| SIDR | 0.039/-0.00097 | 0.14/0.084 | 0.22/0.088 | 0.00 | 77.8 |
| LM | -1.30/-1.16 | 0.097/0.085 | 0.55/0.056 | 0.00 | 44.4 |
| without effect of $X$ on $L$ | | | | | |
| | bias (mean/med) | boot SE (mean/med) | SE (emp/MCD) | p-value bias= 0 | boot cov |
| IPIW | -0.008/0.022 | 2.78/0.033 | 1.25/0.035 | 0.00 | 39.6 |
| DR | -0.17/-0.24 | 0.39/0.11 | 2.34/0.12 | 0.00 | 58.8 |
| UW | -0.34/-0.30 | 0.084/0.060 | 0.15/0.072 | 0.00 | 0.30 |
| SDR | -0.17/-0.22 | 7.2/0.14 | 3.10/0.15 | 0.00 | 73.3 |
| IDR | 0.0003/-0.008 | 0.069/0.056 | 0.10/0.049 | 0.037 | 64.9 |
| SIDR | 0.009/0.009 | 0.081/0.053 | 0.096/0.055 | 0.00 | 54.5 |
| LM | -0.34/-0.30 | 0.024/0.021 | 0.14/0.063 | 0.00 | 21.3 |

Table 6.5: *Results of Simulation Experiment 5*

## 6.4 Data analysis

De Sutter et al. (2006) estimate the effect of single versus double embryo transfer (SET versus DET) on birth weight using a survey of 557 SET and 396 DET patients who entered the subfertility program at the Ghent University hospital and who delivered a singleton child of at least 500 grams after fresh embryo transfer in a first, second or third cycle between January 2003 and May 2007. The mean gestational age (GA) of singleton babies is 273.9 days (SD 12.4). The mean birth weight (BW) is 3231.8 grams (SD 565.4). De Sutter et al. (2006) observed birth weights to be 120 grams (95% confidence interval 44 - 197) lower on average in babies born after double than single enmbryo transfer. In response to criticism that the analysis was not adjusted for gestational age, Delbaere et al.(2007b) argue that such adjustment would remove a possible indirect effect of SET/DET on birth weight through gestational age, and would introduce bias because gestational age may be affected by SET/DET and is associated with birth weight. At the same time, the debate raises the question whether the effect of SET/DET on birth weight is entirely mediated through gestational age.

To address this question, we assume that the causal diagram of Figure 6.1 represents the data generating mechanism, with $S$ representing measured baseline confounders (embryo quality, duration of infertility, maternal age, female and male pathology, gravida and type of conception (IVF/ICSI)) for the association between SET/DET and pregnancy outcomes, and $L$ representing measured confounders (complications during pregnancy, vaginal blood loss, preterm contractions, preterm rupture of the membranes and growth retardation) for the association between gestational age and birth weight. The diagram allows for the presence of unmeasured confounders $U$ for the association between these confounders and outcome $Y$. Note that the analysis is restricted to women who deliver a singleton baby and that an implicit assumption in the analysis is thus that the loss of an embryo (in early pregnancy) in women with DET treatment is not associated with gestational age and birth weight. A more appropriate analysis would be restricted to the principal stratum (Frangakis and Rubin, 2002) of women who would deliver a singleton child (of at least 500 grams) under

both treatments.

Of all variables listed above as potential baseline confounders for the assocation between embryo transfer (SET/DET) and the outcome BW, only maternal age, embryo quality, duration of infertility and IVF/ICSI treatment showed a significant association with SET/DET and/or BW. Thus, only these variables are included as confounders $S$. For similar reasons, only preterm contractions, preterm rupture of the membranes and growth retardation are included as confounders $L$. Due to the many missing values for duration of infertility (33.6%), we first performed the analyses assuming that infertility duration does not confound the association between GA and BW. This leaves us with 895 complete observations.

To estimate the direct effect of SET/DET on BW, which is not mediated by GA, we use the approaches proposed in Section 6.2. Based on the results of the simulation experiments in the previous section, we use the sequential G-estimator with a linear conditional model for $Y$ as the primary estimator in the analysis. Since GA is skewly distributed to the left, we transformed it via a Box-Cox transformation so that we could assume a normal distribution for model (6.8), with mean $\alpha_0 + \alpha_1 L + \alpha_2 S + \alpha_3 X$ and constant residual standard deviation $\sigma_K$. Further, we postulated $m(X, K, S; \psi) = \psi X$, chose $d_K(X, S) = X$ and $q_K(S) = 0$ and we specified linear models $E(X|S; \beta) = \beta_0 + \beta_S$ and $E(Y|K, L, X, S) = \gamma_0 + \gamma_1 K + \gamma_2 L + \gamma_3 X + \gamma_4 S$ for the conditional expectations of the exposure SET/DET ($X$) and the outcome BW ($Y$).

Table 6.6 summarizes the estimates obtained from the different estimation methods, along with bootstrap standard errors and confidence intervals based on 1000 bootstrap samples. As expected, after removing the indirect effect through GA, we now estimate the average birth weight to be merely 60 grams (95% confidence interval 14 - 136) lower on average in babies born after double than single enmbryo transfer. While the difference in birth weight is no longer significant after controlling for GA, the confidence interval does not exclude the possibility of important differences exceeding 100 grams.

| | Without infertility duration | | | With infertility duration | | |
|---|---|---|---|---|---|---|
| | $\hat{\psi}$ | boot SE | 95% CI | $\hat{\psi}$ | boot SE | 95% CI |
| IPIW | 78.20 | 100.16 | [-143.41;304.21] | 91.90 | 144.41 | [-224,78;338.44] |
| DR | -67.77 | 40.44 | [-141.19;14.42] | -84.11 | 53.75 | [-190.64;14.79] |
| UW | -59.64 | 36.70 | [-136.49;13.97] | -70.76 | 47.92 | [-156.37;14.98] |
| SDR | -67.69 | 40.38 | [-141.22;14.38] | -83.82 | 53.54 | [-189.42;15.37] |
| IDR | -69.52 | 42.46 | [-154.40;18.41] | -86.07 | 54.87 | [-181.93;15.03] |
| SIDR | -69.45 | 42.33 | [-153.08;18.18] | -85.77 | 54.59 | [-181.63;14.70] |
| LM | -44.59 | 33.49 | [-115.06;28.88] | -71.14 | 45.18 | [-148.02;6.27] |

Table 6.6: *Data Analysis Results*

## 6.5 Discussion

Estimating the direct effect of an exposure on an outcome, which is not mediated by some given variable, requires adjustment not only for prognostic factors of the outcome that are associated with the exposure, but additionally for those associated with the mediator. In practice, several of these prognostic factors may only arise after the exposure was administered and thus possibly be affected by it. In such settings, standard regression methods may yield biased estimates of the direct exposure effect.

While methods based on inverse probability weighting have been proposed to accommodate this problem, they require inverse weighting by a density when the mediator is discrete with many levels or absolutely continuous. Inefficient effect estimators with large bias are then typically obtained. In this chapter, we have proposed a sequential G-estimator (or, more generally, unweighted estimator) which mitigates this problem by avoiding the inverse weighting altogether. This estimator competes remarkably well with ordinary least squares estimators in settings where these are valid (i.e. in settings where prognostic factors of the outcome which are predictive of the mediator, are not themselves affected by the exposure), but remains valid in settings where the ordinary least squares estimator fails. The proposed estimator requires postulating a working model for the expected outcome in function of exposure, mediator and prognostic fac-

tors. It is robust against misspecification of this working model when the exposure, mediator and its prognostic factors have a multivariate normal distribution, but not otherwise. In view of this, we have derived doubly-robust estimators which allow for misspecification, provided that a working model for a conditional density of the mediator is correctly specified. On the basis of the simulation studies, we recommend the sequential (unweighted) G-estimator and the (stabilized) improved doubly-robust estimator when the mediator is absolutely continuous. For a dichotomous mediator, less variable inverse weights are expected, and thus a relatively much better performance of the (stabilized) improved doubly-robust estimator.

A number of restrictions are implicit in our approach. First, we have implicitly assumed that the mediator may affect the outcome, but is not itself affected by it. In many practical studies, mediator and outcome may mutually affect each other over time. We plan to accommodate this by allowing for repeated measurements on mediator and outcome. Second, we have implicitly assumed that controlled direct effects are well defined. In certain situations however, the idea of fixing the intermediate variable at a value equal for all subjects is not realistic (see Chapter 4 for examples). Standardized direct effects (Didelez, Dawid and Geneletti, 2006) are more broadly useful in the sense that they allow each subject to have their own fixed value for the intermediate variable. Moreover, they can be obtained by averaging the controlled direct-effect estimates in this chapter over a chosen mediator density, under additional assumptions (Petersen, Sinisi and van der Laan, 2006).

# Appendix 6.A1: Proof of Theorem 5

Part 2 of Theorem 5 is immediate upon applying Theorem 1.2 in van der Laan and Robins (2003). To prove Part 1 of Theorem 5, we assume that the regularity conditions of Theorem 1A in Robins, Mark and Newey (1992) hold for $U_{i,DR}(d,q;\psi,\alpha,\beta,\gamma)$, the estimating function $G_i(\gamma)$ for $\gamma$ and $A_i(\alpha)$ for $\alpha$. For simplicity, we assume that $\beta^*$ is known, as is usually the case when $X$ is a randomly assigned exposure. By standard Taylor expansion arguments, we have that

$$\begin{aligned} 0 \quad = \quad & n^{-1/2}\sum_{i=1}^{n} U_{i,DR}(d,q;\psi^*,\tilde{\alpha},\beta,\tilde{\gamma}) \\ & + E\left\{\frac{\partial}{\partial\psi}U_{i,DR}(d,q;\psi=\psi^*,\tilde{\alpha},\beta,\tilde{\gamma})\right\}\sqrt{n}(\widehat{\psi}-\psi^*) \\ & - E\left\{\frac{\partial}{\partial\gamma}U_{i,DR}(d,q;\psi^*,\tilde{\alpha},\beta,\gamma=\tilde{\gamma})\right\}E^{-1}\left\{\frac{\partial}{\partial\gamma}G_i(\gamma=\tilde{\gamma})\right\}G_i(\tilde{\gamma}) \\ & - E\left\{\frac{\partial}{\partial\alpha}U_{i,DR}(d,q;\psi^*,\alpha=\tilde{\alpha},\beta,\tilde{\gamma})\right\}E^{-1}\left\{\frac{\partial}{\partial\alpha}A_i(\alpha=\tilde{\alpha})\right\}A_i(\tilde{\alpha}) + o_p(1) \end{aligned} \tag{6.24}$$

where $o_p(1)$ denotes a random variable converging to 0 in probability, and where $\tilde{\gamma}$ and $\tilde{\alpha}$ are the probability limits of the estimators for $\gamma^*$ and $\alpha^*$.

First note that $U_{i,DR}(d,q;\psi,\tilde{\alpha},\beta,\tilde{\gamma})$ has mean zero at $\psi=\psi^*$ under model $\mathcal{A}$, even when model (6.16) for the conditional expectation of the outcome is misspecified. This is because the first term in (6.13) has mean zero at $\psi^*$ under model $\mathcal{A}$ by construction and the second term is a mean zero function under model $\mathcal{A}$ for each choice of $\phi(K_i,L_i,X_i,S_i)$ and thus in

particular for $\phi_{opt}(K_i, L_i, X_i, S_i)$. We now show that $U_{i,DR}(d, q; \psi, \tilde{\alpha}, \beta, \tilde{\gamma})$ has mean zero at $\psi = \psi^*$ under model $\mathcal{B}$, even when model (6.8) for the conditional density of the mediator is misspecified. Using the potential outcomes framework and denoting $F \equiv f(K|L, S, X; \alpha)$, $F^* \equiv f(K|S)$, $F_k^* \equiv f(K = k|S)$, $M \equiv m(X, S, K; \psi)$, $Q \equiv q_K(S)$ and $D \equiv d_K(X, S)$, we may rewrite the estimating function in (6.18) as

$$\begin{aligned} U_{DR} &= \int \frac{I(K=k)}{F} F^* \Delta\{D|S\}(Y_k - M - Q)dk \\ &\quad -E\left(\int \frac{I(K=k)}{F} F^* \Delta\{D|S\}(Y_k - M - Q)dk|X, K, L, S\right) \\ &\quad +E\left[E\left(\int \frac{I(K=k)}{F} F^* \Delta\{D|S\}(Y_k - M - Q)dk|X, K, L, S\right)|X, L, S\right] \end{aligned}$$

We rewrite the first term as

$$\int \left[F_k^* \Delta\{D|S\}(Y_k - M - Q) + \left\{\frac{I(K=k)}{F} - 1\right\} F_k^* \Delta\{D|S\}(Y_k - M - Q)\right]$$

The second term equals

$$\int \frac{I(K=k)}{F} F^* \Delta\{D|S\} E(Y_k - M - Q|K = k, X, L, S)dk$$

and the third term can be further simplified to

$$\begin{aligned} &E\left[\int \frac{I(K=k)}{F} F^* \Delta\{D|S\} E(Y_k - M - Q|K = k, X, L, S)dk|X, L, S\right] \\ &= \int F_k^* \Delta\{D|S\} E(Y_k - M - Q|K = k, X, L, S) E\left(\frac{I(K=k)}{F}|X, L, S\right) dk \\ &= \int F_k^* \Delta\{D|S\} E(Y_k - M - Q|K = k, X, L, S)dk \end{aligned}$$

Adding these 3 terms yields

$$\begin{aligned} &\int \left[F_k^* \Delta\{D|S\}(Y_k - M - Q) + \left\{\frac{I(K=k)}{F} - 1\right\} F_k^* \Delta\{D|S\}(Y_k - M - Q)\right. \\ &\left. - \left\{\frac{I(K=k)}{F} - 1\right\} F_k^* \Delta\{D|S\} E(Y_k - M - Q|K = k, X, L, S)\right] dk \\ &= \int (Y_k - M - Q) F_k^* \Delta\{D|S\} dk \\ &+ \int \left\{\frac{I(K=k)}{F} - 1\right\} F_k^* \Delta\{D|S\}(Y_k - E(Y_k|K = k, X, L, S))dk \end{aligned}$$

The first term was shown to have mean zero at $\psi^*$ in Section 6.2.2. The integrand of the second term has mean zero conditional on $(K, X, L, S)$ when, as in model $\mathcal{B}$, the conditional expectation of $Y$ is correctly specified, since $Y_k$ is independent of $K$ conditionally on $X$, $L$ and $S$.

We conclude that $U_{i,DR}(d, q; \psi, \tilde{\alpha}, \beta, \tilde{\gamma})$ has mean zero at $\psi = \psi^*$ under model $\mathcal{A} \cup \mathcal{B}$. Further, note that $\tilde{\gamma} = \gamma^*$, and thus that $E\left\{\partial U_{i,DR}(d, q; \psi^*, \alpha = \tilde{\alpha}, \beta, \tilde{\gamma})/\partial\alpha\right\} = 0$ and $E\{G_i(\tilde{\gamma})\} = 0$ when model (6.16) is correctly specified. Likewise, $\tilde{\alpha} = \alpha^*$, and thus $E\left\{\partial U_{i,DR}(d, q; \psi^*, \tilde{\alpha}, \beta, \gamma = \tilde{\gamma})/\partial\gamma\right\} = 0$ and $E\{A_i(\tilde{\alpha})\} = 0$ when model (6.8) is correctly specified. Under the regularity conditions of Theorem 1A in Robins, Mark and Newey (1992), it now follows from the asymptotic unbiasedness of $\sqrt{n}(\widehat{\psi} - \psi^*)$ under model $\mathcal{A} \cup \mathcal{B}$ that $\widehat{\psi}$ is a consistent estimator of $\psi^*$ under model $\mathcal{A} \cup \mathcal{B}$.

# Discussion

The main challenge in causal inference is protecting estimators of the effect of an exposure on an outcome against possible confounders. This is important because improper or insufficient adjustment for confounders may yield estimated effects that reflect merely associations in the data and not causal effects. In particular, such associations may be found even in the absence of a causal effect. In this thesis, we were motivated by research questions concerning twin data, infertility and perinatal outcomes that were raised through collaborations with the Department of Obstetrics and Gynecology. This has lead to the development of new methods for inferring causal effects.

In the first part of the thesis, we considered the use of twin data for estimating heritability. We learned that estimates of heritability may be confounded, in the sense that standard estimates obtained using structural equation models may suggest a trait to be heritable even when it is not. This may happen, even when the equal environment assumption appears to hold, whenever there exist common causes of zygosity and the trait of interest. Causal directed acyclic graphs (Pearl, 2000) help gain insight into this and suggest that problems of confounding may not be very common when estimating heritability.

Next, we evaluated how to estimate the effect of an exposure (e.g. smoking) on a given outcome (e.g. lung function) based on twin data. For such studies, it has been suggested, on the basis of data analysis and simulation evidence, that the separation of exposures into within- and between-cluster components within a random intercept model yields protection against unmeasured twin-specific confounders (Neuhaus and Kalbfleisch, 1998; Car-

lin, 2005). We investigated the validity of these methods by first developing a general class of conditional generalized estimating equations for the estimation of causal effects from general clustered data (e.g. data from multi-center studies, family data, longitudinal data, ...), which offer protection against confounders that have a constant value within each cluster (cluster-level confounders). This is done by exploiting the correlated structure of the data and making comparisons within clusters. Next, we evaluated whether estimators obtained by separating exposures into within- and between-cluster components can be viewed as estimators within our class. We expressed some concerns over these estimators because (a) models which involve within- and between-cluster exposures cannot be viewed as data-generating models; and (b) such approaches may be inconsistent and inefficient under nonlinear link functions. Nonetheless, we show the latter approaches to be very useful and attractive because they offer computationally convenient estimators, they are valid under the identity link and approximately valid, but inefficient, under the log link.

Conditional generalized estimating equations were developed to obtain protection against unmeasured cluster-level confounders. In the context of longitudinal studies, that is, to offer protection against unmeasured baseline confounders. Future research would be useful to extend these methods to marginal structural models (Robins et al., 2000; Yu and van der Laan, 2006) or structural nested mean models (Robins, 1999b) for the effect of time-varying exposures in longitudinal studies.

From Chapter 4 onwards, we focused on estimation of direct causal effects. This was motivated by research questions at the Department of Obstetrics and Gynecology on the effect of subfertility treatments on perinatal health, which is not mediated by the beneficial effect through zygosity. Direct effects play an important role in many fields of research. This is because exposures often affect the outcome through various pathways, both indirectly through intermediate variables and directly. Understanding and estimating the different path-specific effects is then important to gain insight into the mechanistic action of a treatment (e.g. a drug) or intervention or to evaluate the importance of specific components of an intervention For example, Petersen, Sinisi and van der Laan (2006) estimate the effect of

protease inhibator-based antiretroviral therapy on CD4 T-cell count (an indication for HIV infection). They then investigate whether this beneficial effect is entirely due to a reduction in plasma HIV RNA level (viral load). In randomized clinical studies for the effect of postmenopausal hormone therapy on breast cancer, women in the treated group tend to undergo a mammography more often which leads to a more early detection of breast cancer than in the untreated group, and thus to better survival chances (Gajdos et al., 2000). It is then of interest to investigate whether the treatment directly affects the survival chances for breast cancer patients, not mediated through mammography. Direct effects are also relevant in many non-medical contexts. In the context of company management, direct effects are of interest when companies make decisions about their marketing strategy and wish to know the impact on the behavior of customers. These decisions also have an indirect effect, which is not of interest, since they affect the reaction of competitive companies which also influences the customer. In human resources, questions of race or gender discrimination are frequently posed when hiring people. To examine this, one needs to acknowledge that race and gender may affect the educational level, career objectives, etc., which in turn affects the decision to hire someone. This indirect effect through educational level, career objectives, etc., does not reflect race or gender discrimination and thus only the direct effect is of interest.

Estimation of direct effects is more complex than estimation of (total) causal effects. This is so because estimation of direct effects requires adjustment for intermediate variables, which may induce non-causal relations between the exposure and the outcome when, as is likely the case in practice, prognostic factors of the outcome also affect the intermediate variable. Even when some of these prognostic factors are measured and can be adjusted for, standard adjustment remains problematic when the exposure affects these factors and when there are unmeasured confounders for these factors and the outcome. Robins introduced inverse probability weighting for structural nested direct effects models (1999b) to infer controlled direct effects (i.e. the effect of an exposure on an outcome while holding the intermediate variable fixed at a certain value for all subjects). This is based

on inversely weighting each subject's data by a conditional distribution of the intermediate variable. When the latter is absolutely continuous, these weights may be very unstable and lead to extremely imprecise direct effect estimates. In this thesis, we have therefore developed doubly robust estimators. These are consistent when either the model for the weights is correctly specified or a conditional mean model for the outcome. By purposely misspecifying the weights to equal 1, the weight instability is avoided and consistent and efficient estimators are obtained provided that the conditional mean model for the outcome is correctly specified. To make the resulting unweighted estimator more robust against model misspecification, we additionally developed doubly robust estimators that cope better with weight instability.

Further research is of interest to extend these methods to longitudinal data. While methods for longitudinal data have been proposed (Rosenblum et al., 2007), they are based on inverse probability weighting and thus also fail to work well when the intermediate variable is continuous. It is hence of interest to develop approaches that work well even with an absolutely continuous intermediate variable.

A further extension of interest is to estimate natural direct effects. These represent the effect of exposure on outcome when removing the exposure effect on the intermediate variable. Petersen, Sinisi and van der Laan (2006) proposed natural direct effects estimators, but only for cases where confounders for the association between intermediate and outcome are not themselves affected by the exposure. Allowing for greater flexibility is of interest.

We have applied the unweighted estimator to estimate a direct genetic effect on lung function, not mediated through body mass. Our method so far assumes that there are no ascertainment conditions (i.e., it assumes having a random sample of data). However, many genetic studies sample study subjects on the basis of their outcome. Vansteelandt et al. (2007) show in the context of family-based designs that the ascertainment is not problematic when there is no (total) genetic effect (i.e. the combination of direct and indirect effects trough intermediate variables). In particular, tests for a genetic effect will not incorrectly reject the null hypothesis more often

than a priori stated. The reason is that under no (total) genetic effect, the SNPs and the potential outcome are independent (i.e. d-separated) within the group of ascertained subjects. A challenge would be to develop direct effect tests which are valid in the presence of ascertainment conditions.

# Bibliography


Abadie, A. (2005). Semiparametric difference-in-differences estimators. *Review of Economic Studies* **72**, 1-19.

Albert, J.M. (2002). Estimating efficacy in clinical trials with clustered binary responses. *Statistics in Medicine* **21**(5), 649-661.

Baron, M.B., Kenny, D.A. (1986). The moderator-mediator variable distinction in social psychological research: conceptual, strategic, and statistical considerations. *Journal of Personality and Social Psychology* **51**(6), 1173-1183.

Barret-Connor, E., Grady, D. (1998). Hormone replacement therapy, heart disease, and other considerations. *Annual Review of Public Health* **19**, 55-72.

Begg, M.D., Parides, M.K. (2003). Separation of individual-level and cluster-level covariate effects in regression analysis of correlated data. *Statistics in Medicine* **22**, 2591-2602.

Berlin, J.A., Kimmel, S.E., Ten Have, T.R., Sammel, M.D. (1999). An empirical comparison of several clustered data approaches under confounding due to cluster effects in the analysis of complications of coronary angioplasty. *Biometrics* **55**, 470-476.

Bickel, P.J., Hammel, E. A., O'Connell, J. W. (1975). Sex Bias in Graduate Admissions: Data from Berkeley. *Science* **187**, 398-404.

Bickel P.J., Klaassen C.A.J., Ritov Y., Wellner J.A. (1993). *Efficient and adaptive estimation for semiparametric models.* New-York : Springer-Verlag.

Bierut, L., Madden, P., Breslau, N., Johnson, E., Hatsukami, D., Pomerleau, O., Swan, G., Rutter, J., Bertelsen, S., Fox, L., et al. (2007). Novel genes identified in a high-density genome wide association study for nicotine dependence. *Human Molecular Genetics* **16**(1), 24.

Blondel, B., Kaminski, M. (2002). Trends in the occurrence, determinants, and consequences of multiple births. *Semin Perinatol* **26**,239-49.

Blondel, B., Macfarlane, A. (2003). Rising multiple maternity rates and medical management of subfertility: better information is needed. *European Journal of Public Health* **13**,83-6.

Camargo Jr, C., Weiss, S., Zhang, S., Willett, W., Speizer, F. (1999). Prospective study of body mass index, weight change and risk of adult-onset asthma in women. *Archives of internal medicine* **159**(21), 2582.

Carlin, J.B., Gurrin,L.C., Sterne, J.A.C, Morley, R., Dwyer, T. (2005). Regression models for twin studies: a critical review. *International Journal of Epidemiology* **34**, 1089-1099.

Carmelli, D., Page, W.F. (1996). Twenty-four mortality in World War II US veteran twins discordant for cigarette smoking. *International Journal of Epidemiology* **25**(3), 554-559.

Carmichael, C., McGue, M. (1995). A cross-sectional examiniation of height, weight, and body mass index in adult twins. *Journals of Gerontology Series A: Biological and Medical Sciences* **50**(4), 237-244.

Chao, W.H., Palta, M., Young, T. (1997). Effect of omitted confounders on the analysis of correlated binary data. *Biometrics* **53**, 678-689.

Cole, S.R., Hernan, M.A. (2002). Fallibility in estimating direct effects. *International Journal of Epidemiology* **31**, 163-165.

Dallal, E.G. (2001). *The Little Handbook of Statistical Practice*. Jean Mayer USDA Human Nutrition Research Center on Aging at Tufts University, 711 Washington Street, Boston. http://www.tufts.edu/ gdallal/cause.htm

Dawid, A.P. (1979). Conditional independence in statistical theory. *Journal of Royal Statistical Society, Series A* **41**, 1-31.

Delbaere, I., Verstraelen, H., Goetgeluk, S., Martens, G., De Backer, G., Temmerman, M. (2007a). Pregnancy outcome in primiparae of advanced maternal age . *European Journal of Obstetrics, Gynecology and Reproductive Biology* **135**, 41-46.

Delbaere, I., Vansteelandt, S., De Bacquer, D., Verstraelen, H., Gerris, J., De Sutter, P., Temmerman, M. (2007b). 'Should we adjust for gestational age when analysing birth weights? The use of z-scores revisited. *Human Reproduction* **22**, 2080-2083.

Demissie, K., Brechenridge, M., Rhoads, G. (1998). Infant and maternal outcomes in the pregnancies of asthmatic women. *American Journal of Respiratory and Critical Care Medicine* **158**(4), 1091-1095.

Derom, C., Derom, R. (2005). The East Flanders Prospective Twin Survey. In: Blickstein I, Keith LG, eds. *Multiple pregnancy: epidemiology, gestation and perinatal outcome*. 2nd ed. Oxford: Taylor and Francis, 39-47.

De Sutter, P., Delbaere, I., Gerris, J., Verstraelen, H., Goetgeluk, S., Van der Elst, J., Temmerman, M., Dhont, M. (2006). Birthweight of singletons after assisted reproduction is higher after single- than after double-embryo transfer. *Human Reproduction* **21**, 2633-2637.

Devlin, B., Roeder, K. (1999). Genomic control for association studies. *Biometrics*. **55**, 997-1004.

Didelez, V., Dawid, A.P., Geneletti, S. (2006). Direct and Indirect Effects of Sequential Treatments. *Proceedings of the 22nd Annual Conference on Uncertainty in Artifical Intelligence*, 138-146

Diggle, P., Liang, K.Y., Zeger, S.L. (1994). *Analysis of Longitudinal Data*. Oxford University Press: Oxford.

Doll, R., Hill, A.B. (1954). The mortality of doctors in relation to their smoking habits - A preliminary report. *Britisch Medical Journal* **2**, 1451-1455.

Epstein, M., Allen, A., Satten, G. (2007). A simple and improved correction for population stratification in case-control studies. *American Journal of Human Genetics* **80**(5), 921-930.

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

Frayling, T., Timpson, N., Weedon, M., Zeggini, E., Freathy, R., Lindgren, C., Perry, J., Elliott, K., Lango, H., Rayner, N., et al. (2007). A common variant in the FTO gene is associated with body mass index and predisposes to childhood and adult obesity. *Science* **316**(5826), 889.

Gajdos, C., Tartter, P.I., Babinszki, A. (2000). Breast Cancer Diagnosed During Hormone Replacement Therapy. *Obstetrics and Gynecology* **95**, 513-518.

Gessner, B., Chimonas, M. (2007). Asthma is associated with preterm birth but not with small for gestational age status among a population-based cohort of Medicaid-enrolled children; 10 years ago. *British Medical Journal* **62**(3), 231-236.

Goetgeluk, S., Vansteelandt, S. (2008). Conditional generalized estimating equations for the analysis of clustered and longitudinal data. *Biometrics* doi:10.1111/j.1541-0420.2007.00944.x

Goetgeluk, S., Vansteelandt, S., Goetghebeur, E. (2008). Estimation of controlled direct effects. *Harvard University Biostatistics Working Paper Series*. Berkeley Electronic Press.

Goetghebeur, E., Lapp, K. (1997). The Effect of Treatment Compliance in a Placebo-Controlled Trial: Regression with Unpaired Data. *Applied Statistics* **46**, 351-364.

Greenland, S., Pearl, J., Robins, J.M. (1999). Causal diagrams for epidemiologic research. *Epidemiology* **10**(1), 37-48.

Greenland, S., Brumback, B. (2002). An overview of relations among causal modelling methods. *International Journal of Epidemiology* **31**(5), 1030-1037.

Group C.A.M.P. (1999). The childhood asthma management program (camp): design, rationale, and methods. *Controlled clinical Trials* **20**, 91-120.

Guay, L.A., Musoke, P., Fleming, T., Bagenda, D., Allen, M., Nakabiito, C., Sherman, J., Bakaki, P., Ducar, C., Deseyve, M., Emel, L., Mirochnick, M., Fowler, M.G., Mofenson, L., Miotti, P., Dransfield, K., Bray, D., Mmiro, F., Jackson, J.B. (1999). Intrapartum and neonatal single-dose nevirapine compared with zidovudine for prevention of mother-to-child transmission of HIV-1 in Kampala, Uganda: HIVNET 012 randomised trial. *The Lancet* **354**(9181),795-802.

Heckman, J.J. (1979). Sample selection bias as a specification error. *Econometrica* **47**, 153-161.

Helmerhorst, F.M., Perquin, D.A., Donker, D., Keirse, M.J. (2004). Perinatal outcome of singletons and twins after assisted conception: a systematic review of controlled studies. *BMJ* **328**, 261.

Herbert, A., Gerry, N., McQueen, M., Heid, I., Pfeufer, A., Illig, T., Wichmann, E.-H., Meitinger, T., Hunter, D., Hu, F., Colditz, G., Zhu, X., Cooper, R., Ardlie, K., Lyon, H., Hirschhorn, J., Laird, N, Lenburg, M., Lange, C., Christman, M. (2006). Genetic variation near insig2 is a common determinant of obesity in western europeans and african americans. *Science* **312**(5771), 279-283.

Hernan, M.A., Hernandez-Diaz, S., Weler, M.M., Mitchell, A.A. (2002). Causal knowledge as a prerequisite for confounding evaluation: an application to birth defects epidemiology. *American Journal of Epidemiology* **155** (2): 176-184.

Hernan, M.A. (2004). A definition of causal effect for epidemiological research. *Journal for epidemiology and community helath* **58**(4), 265-271.

Hernan, M.A., Cole, S.R. (2002). Fallibility in estimating direct effects. *International Journal of Epidemiology* **31**, 163-166.

Hernan, M.A., Hernandez-Diaz, S., Robins, J.M. (2004). A structural approach to selection bias. *Epidemiology* **15**, 615-625.

Hirano, K., Imbens, G.W., Rubin, D.B., Zhou, X.H. (2000). Assessing the effect of an influenza vaccine in an encouragement design. *Biostatistics* **1**, 69-88.

Hirschhorn, J., Lindgren, C., Daly, M., Kirby, A., Schaffner, S., Burtt, N., Altshuler, D., Parker, A., Rioux, J., Platko, J., et al. (2001). Genomewide linkage analysis of stature in multiple populations reveals several regions with evidence of linkage to adult height. *the American Journal of Human Genetics* **69**(1), 106-116.

Hoekstra, C., Zhao, Z.Z., Lambalk, C.B., Willemsen, G., Martin, N.G., Boomsma, D.I., Montgomery, G.W. (2008). Dizygotic twinning. *Human Reproduction* **14**(1), 37-47.

Ionita-Laza, I., McQueen, M., Laird, N., Lange, C. (2007). Genomewide weighted hypothesis testing in family-based association studies, with an application to a 100k scan. *American Journal of Human Genetics* **81**, 607-614.

Jackson, R.A., Gibson, K.A., Wu, Y.W., Croughan, M.S. (2004). Perinatal outcomes in singletons following in vitro fertilization: a meta-analysis. *Obstetrics in Gynecology* **103**,551-63.

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

Koziel, S, Ulijaszek SJ, Szklarska A, Bielicki T. (2007). The effects of fatness and fat distribution on respiratory functions *Annals of Human Biology* **34**(1), 123-131.

Laird, N.M., Lange C. (2006). Family-based designs in the age of large-scale gene-association studies. *Nature Reviews Genetics* **7**(5), 385-394.

Laird, N., Horvath, S., Xu, X. (2000). Implementing a unified approach to family based tests of association. *Genetical Epidemiology* **19(Suppl 1)** (S36-S42).

Laird, N.M., Ware, J.H. (1985). Random-effects models for longitudinal data. *Biometrics* **38**(4), 963-974.

Lake, S.L., Blacker, D., Laird, N.M. (2000). Family-based tests of association in the presence of linkage. *Am J Hum Genet* **67**, 15151525.: A powerful new testing strategy. *American Journal of Human Genetics* **79**, 801-811.

Lange, C., DeMeo, D., SIlverman, E., Weiss, S., Laird, N. (2003). Using the noninformative families in family-based association tests

Li, M., Kane, J., Konu, O. (2003). Nicotine, body weight and potential implications in the treatment of obesity. *Curr Top Med Chem* **3**(8), 899-919.

Liang, K.Y., Zeger, S.L. (1986). Longitudinal data analysis using generalized linear models. *Biometrika* **73**, 13-22.

Loeys, T., Vansteelandt, S., Goetghebeur, E. (2001). Accounting for correlation and compliance in cluster randomized trials. *Statistics in Medicine* **20** (24), 3753 - 3767.

Loos, R., Derom, C., Vlietinck, R., Derom, R. (1998). The East Flanders Prospective Twin Survey (Belgium): a population-based register. *Twin Research* **1**,167-75.

Louis, T.A., Robins, J., Dockery, D.W., Spiro, A., Ware, J.H. (1986). Explaining discrepancies between longitudinal and cross-sectional models. *Journal of Chronic Diseases* **39**, 831-893.

Lyon H., Lange C., Lake S., Silverman E.E., Randolph A.G., Kwiatkowski D., Raby B.A., Lazarus R., Weiland K.M., Laird N., Weiss S.T. 2004. IL10 gene polymorphisms are associated with asthma phenotypes in children. *Genetic Epidemiology* **26**(2), 155-165.

Machin, G.A. (2004). Why is it important to diagnose chorionicity and how do we do it? *Best Practical and Research: Clinical Obstetrics and Gynaecology* **18**,515-530.

Mackinnon, D.P., Lockwood, C.M., Hoffman, J.M., West, S.G., Sheets, V. (2002). A comparison of methods to test mediation and other intervening variables effects. *Psychological Methods* **7**(1), 83-104.

Mark, S.D., Robins, J.M. (1993). Estimating the causal effect of smoking cessation in the presence of confounding factors using a rank preserving structural failure time model. *Statistics in Medicine* **12**(17), 1605-1628.

Manson, J.E.,Hsia, J., Johnson, K.C. et al. (Women's Hlth Initiative Investigat) (2003). *National England Journal of Medicine* **349**, 523-534.

Molenberghs, G., Burzykowski, T., Alonso, A., Buyse, M. (2004). A perspective on surrogate endpoints in controlled clinical trials. *Statistical Methods in Medical Research* **13**, 177-206.

Neale, M.C., Cardon, L.R. (1992). *Methodology for Genetic Studies of Twins and Families*. Kluwer Academic Publishers, The Netherlands.

Neuhaus, J.M., Kalbfleisch, J.D. (1998). Between- and within-cluster covariate effects in the analysis of clustered data. *Biometrics* **54**, 638-645.

Neuhaus, J.M., McCulloch, C.E. (2006). Separating between- and within-cluster covariate effects by using conditional and partitioning methods. *Journal of the Royal Statistical Society Series B* **68**, 859-872.

Newey, W.K., McFadden, D. (1994). Large sample estimation and hypothesis testing. In: Engle, R., McFadden, D. (Eds.), *The Handbook of Econometrics*, vol. 4. North-Holland, Amsterdam.

Oberle, D., von Mutius, E., van Kries, R. (2003). Childhood asthma and continuous exposure to cats since the first year of life with cats allowed in the child's bedroom. *Allergy* **58**, 1033-1036.

Oliveti, J.F., Kercsmar, C.M., Redline, S. (2006). Pre-and Perinatal Risk Factors for Asthma in Inner City African-American Children. *American Journal of Epidemiology* **143**(6), 570-577.

Palta, M., Yao, T.J. (1991). Analysis of longitudinal data with unmeasured confounders. *Biometrics* **47**, 1335-1369.

Pearl J. (1995). Causal diagrams for empirical research. *Biometrika* **82**(4), 669-688.

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

Pearl, J. (2001). Direct and Indirect Effects. In *Proceedings of the Seventeenth Conference on Uncertainty in Artificial Intelligence*, ed. M. Kaufmann, San Francisco, CA, 411-420.

Pearson, K., Lee, A., Bramley-Moore, L. (1899). Mathematical contributions to the theory of evolution: VI Genetic (reproductive) selection: Inheritance of fertility in man, and of fecundity in thoroughbred racehorses. *Philosophical Transactions of the Royal Society A* **192**,257330.

Permutt, T., Hebel, J.R. (1989). Simultaneous equation estimation in a clinical trial of the effect of smoking on birth weight. *Biometrics* **45**(2), 619-622.

Petersen, M.L., Sinisi, S.E., van der Laan, M.J. (2006). Estimation of direct causal effects. *Epidemiology* **17**, 276-284.

Price, A., Patterson, N., Plenge, R., Weinblatt, M., Shadick, N., Reich, D. (2006). Principal components analysis corrects for stratification in genome-wide association studies. *Nature Genetics* **38**, 904-909.

Pritchard, J.K., Rosenberg, N.A. (1999). Use of unlinked genetic markers to detect population stratification in association studies. *American Journal of Human Genetics* **65**, 220-8.

Quaghebeur, A., Mutunga, L., Mwanyumba, F., Mandaliya, K., Verhofstede, C., Temmerman, M. (2004).Low efficacy of nevirapine (HIVNET012) in preventing perinatal HIV-1 transmission in a real-life situation. *AIDS* **18**(13),1854-1856.

Rabinowitz D., Laird N. (2000). A unified approach to adjusting association tests for population admixture with arbitrary pedigree structure and arbitrary missing marker information. *Human Heredity* **50**(4), 211-223.

Reisby N., Gram L.F., Bech P., Nagy A., Petersen G.O., Ortmann J., Ibsen I., Dencker S.J., Jacobsen O., Krautwald O., Sondergaard I., Christiansen J. (1977). Imipramine: Clinical effects and pharmacokinetic variability. *Psychopharmacology* **54**(3), 263-272.

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

Robins, J.M. (1994). Correcting for non-compliance in randomized trials using structural nested mean models. *Communications in Statistics* **23**, 2379-2412.

Robins JM. (1999a). Marginal Structural Models versus Structural Nested Models as Tools for Causal Inference. In *S*tatistical Models in Epidemiology: The Environment and Clinical Trials, Eds. M.E. Halloran and D. Berry. IMA Volume 116, NY: Springer-Verlag, pp. 95-134.

Robins, J.M. (1999b). Testing and estimation of direct effects by reparameterizing directed acyclic graphs with structural nested models. In *Computation, Causation, and Discovery*, eds. C. Glymour, and G.F. Cooper, AAAI Press/The MIT Press. 349-405.

Robins, J.M. (2001). Data, design, and background knowledge in etiologic inference. *Epidemiology* **12**(3), 313-320.

Robins, J.M., Hernan, M.a., Brumback, B. (2000). Marginal structural models and causal inference in epidemiology. *Epidemiology* **11**(5), 550-560.

Robins, J.M., Greenland, S. (1992). Identifiability and exchangeability for direct and indirect effects. *Epidemiology* **3**, 143-155.

Robins, J.M., Mark, S.D., Newey, W.K. (1992). Estimating exposure effects by modeling the expectation of exposure conditional on confounders. *Biometrics* **48**(2), 479-495.

Robins, J.M., Rotnitzky, A., Vansteelandt, S. (2007). Discussion of 'Principal stratification designs to estimate input data missing due to death' by C.E. Frangakis, D.B. Rubin, M.-W. An, and E. MacKenzie'. *Biometrics* **63**, 650-653.

Robins, J.M., Rotnitzky, A. and Zhao, D. (1994). Estimation of regression-coefficients when some regressors are not always observed. *J*ournal of the American Statistical Association, **8**9, 846-866.

Robins J.M., Smoller J.W., Lunetta K. (2001). On the validity of the TDT test in the presence of comorbidity and ascertainment bias. *Genetic Epidemiology* **21** (4), 326-36.

Robins, J.M., Sued, M., Lei-Gomez, Q., Rotnitzky, A. (2007). Performance of double-robust estimators when 'inverse probability weights are highly variable. *Statistical Science*, in press.

Robins, J.M., Tsiatis, A.A. (1991). Correcting for noncompliance in randomized trials using rank preserving structural failure time models. *Communications in statistics - theory and methods* **20**, 2609-2631.

Rosenbaum, P.R. (1984). The consequences of adjustment for a concomitant variable that has been affected by the treatment. *Journal of the Royal Statistical Society Series A* **147**, 656-666.

Rosenblum, M., Shiboski, S., Jewell, N.P., van der Straten, A., van der Laan, M.J., Padian, N. (2007). Analyzing direct effects in randomized trials with secondary interventions. *U.C. Berkeley Division of Biostatistics Working Papers* **paper 223**.

Rossouw, J.E., Prentice, R.L., Manson, J.E., Wu, L., Barad, D., Barnabei, V.M., Ko, M., LaCroix, A.Z., Margolis, K.L., Stefanick, M.L. (2007). Postmenopausal hormone therapy and risk of cardiovascular disease by age and years since menopause. *Journal of womens health* **16**(6), 927-928.

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

Rubin, D.B. (1978). Bayesian inference for causal effects - Role of randomization. *Annals of Statistics* **6**(1), 34-58.

Rubin, D.B. (2004). Direct and indirect causal effects via potential outcomes. *Scandinavian Journal of Statistics* **31**, 161-170.

Shadt, E., Lamb, J., Yang, X., Zhu, J., Edwards, S., Guhathakurta, D.,Sieberts, S., Monks, S., Reitman, M., Zhang, C., et al. (2005). An integrative genomics approach to infer causal associations between gene expressions and disease. *Nature Genetics* **37**, 710-117.

Sham, P. (1998). *Statistics in Human Genetics*. Arnold Applications of Statistics.

Sin, D., Spier, S., Svenson, L., Schopflocher, D., Senthilselvan, A., Cowie, R., Man, S. (2004). The relationship between birth weight and childhood asthma: a population-based cohort study. *Archives of Pediatrics and Adolescent Medicine* **158**(1), 60-64.

Smoller, J., Lunetta, K., Robins, J. (2000). Implication of comorbidity and ascertainment bias for identifying disease genes. *American Journal of Medical Genetics (Neuropsychiatric Genetics)* **96**, 817-822.

Sommer, A., Tarwotjo, I., Djunaedi, E., West, K.P., Loedin, A.A., Tilden, R., Mele, L. (1986). Impact of vitamin A supplementation on childhood mortality: a randomized controlled community trial. *Lancet* **i**, 1169-1173.

Sommer, I.E.C., Ramsey, N.F., Bouma, A., Kahn, R.S. (1999). Cerebral mirror-imaging in a monozygotic twin. *The Lancet* **355**,1644-1645.

Tan, Z.Q. (2006). A distributional approach for causal inference using propensity scores. *Journal of the American Statistical Association 101*, 1619-1637.

Tavernas, E., Camargo, C., Rifas-shiman, S., Oken, E., Gold, D., Weiss, S., Gillman, M. (2006). Association of birth weight with asthma-related outcomes at age 2 years. *Pediatic Pulmonol* **41**(7), 643-648.

Taylor, A. (2003). ABC of subfertility: extent of the problem. *BMJ* **327**,434-6.

Taylor, J.M.G., Wang, Y., Thiebaut, R. (2005). Counterfactual links to the proportion of treatment effect explained by a surrogate marker. *Biometrics* **61**, 1102-1111.

Ten Have, T.R., Ratcliffe, S.J., Reboussin, B.A., Miller, M.E. (2004). Deviations from the population-averaged versus cluster-specific relationship for clustered binary data. *Statistical Methods in Medical Research* **13**, 3-16.

Ten Have, T.R., Joffe, M.M., Lynch, K.G., Brown, G.K., Maisto, S.A., Beck, A.T. (2007). Causal Mediation Analyses with Rank Preserving Models. *Biometrics* **63**, 926-934.

van der Laan, M. J., Robins, J. M. (2003), *Unified Methods for Censored Longitudinal Data and Causality*. New-York: Springer-Verlag.

van der Laan, M.J., Petersen, M.L. (2004). Estimation of direct and indirect causal effects in longitudinal studies. *U.C. Berkeley Division of Biostatistics Working Paper Series*, **paper 155**.

van der Laan, M.J., Petersen, M.L. (2005). Direct Effect Models, *U.C. Berkeley Division of Biostatistics Working Paper Series*, **paper 187**.

Vansteelandt, S. (2007). On confounding, prediction and efficiency in the analysis of longitudinal and cross-sectional clustered data. *Scandinavian Journal of Statistics* **34**(3), 478-498.

Vansteelandt, S., Goetgeluk S., Waldman, I., Lyon, H., Schadt, E.E., Weiss, S.T., Lange, C. (2008). A general principle for the identification of direct causal genetic pathways in association studies. *under review for publication in American Journal of Human Genetics*.

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

Vansteelandt, S., Goetghebeur, E., Kenward, M.G., Molenberghs, G. (2006). Ignorance and uncertainty regions as inferential tools in a sensitivity analysis. *Statistica Sinica* **16**(3), 953-979.

Vansteelandt S., Demeo D.L., Su J., Smoller J., Murphy A.J., McQueen M., Schneiter K., Celedon J.C., Weiss S.T., Silverman E.K., Lange C. (2007). Testing and estimating gene-environment interactions in family-based association studies.

Verbeke, G., Molenberghs, G. (1997). *Linear Mixed Models in Practice: A SAS-Oriented Approach*. New York : Springer-Verlag.

Verbeke, G., Molenberghs, G. (2000). *Linear Mixed Models for Longitudinal Data*. New York : Springer-Verlag.

Verbeke, G., Spiessens, B., Lesaffre, E. (2001). Conditional linear mixed models. *The American Statistician* **55**, 25-34.

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

Verstraelen, H., Goetgeluk, S., Derom, C., Vansteelandt, S., Derom, R., Goetghebeur, E., Temmerman, M. (2005). Preterm birth in twins following subfertility treatment: a population-based cohort study. *British Medical Journal* **331**, 1173-1176.

Wang, Y., Petersen, M.L., Bangsberg, D., van der Laan, M.J. (2008). Diagnosing Bias in the Inverse Probability of Treatment Weighted Estimator Resulting from Violation of Experimental Treatment Assignment. *The Berkeley Electronic Press*, Working Paper 211.

Yu, Z., van der Laan, M. (2006). Double Robust Estimation in Longitudinal Marginal Structural Models. *Journal of Statistical Planning and Inference* **136**(3), 1061-1089.

Yuan, W., Basso, O., Sorenson, H., Olsen, J. (2002). Fetal growth and hospitalization with asthma during early childhood: a follow-up study in Denmark. REF

Zimlichman, E., Kochba, I., Mimouni, F., Shochat, T., Grotto, I., Kreiss, Y., Mandel, D. (2005). Smoking habits and obesity in young adults. *Addiction* **100**(7), 1021-1025.

# Nederlandse samenvatting

## Situering en probleemstelling

Oorzaak-effectrelaties detecteren en kwantificeren vormt de basis voor procesbeheersing en interventie in wetenschap en industrie. Data worden daartoe op grote schaal vergaard via experimentele of observationele studies en statistische technieken worden aangewend om deze data te analyseren. Echter, standaard statistische technieken zijn gericht op het vinden van associaties tussen metingen en die kunnen bestaan zelfs als er geen causaal verband is en omgekeerd. Dit is het geval wanneer de oorzakelijke factor samengaat met andere variabelen die een eigen impact op de uitkomst hebben. Dergelijke variabelen worden 'confounders' genoemd. Statistische modelbouw kan soms corrigeren voor gemeten confounders, maar niet voor ongekende confounders waarvan men veelal het bestaan vermoedt. Het zoeken naar onvertekende, efficiënte en robuuste methoden voor causale besluitvorming is daarom een belangrijk, maar niettemin zeer complex onderdeel van het statistisch onderzoek.

De laatste 3 decennia werden belangrijke nieuwe inzichten verworven omtrent besluitvorming voor causale effecten (Rubin, 1978; Robins, 1986; Pearl, 1995, 2000). Zo werden potentiële uitkomsten (Rubin, 1978; Robins, 1986) geïntroduceerd die een duidelijke definitie van het causaal effect toelaat, en werden causale diagrammen (Pearl, 1995, 2000) ontwikkeld die het mogelijk maken deze effecten visueel voor te stellen. Een introductie tot causale besluitvorming, voornamelijk omtrent potentiële uitkomsten en causale diagrammen, wordt gegeven in Hoofdstuk 1.

In deze thesis worden causale methodes toegepast en ontwikkeld die algemeen noodzakelijk zijn om enkele specifieke causale problemen aan te pakken, gemotiveerd door onderzoeksvragen vanuit de Vakgroep Uro-Gynaecologie van het Universitair Ziekenhuis Gent en het Department of Biostatistics van de Harvard School of Public Health, betreffende de analyse van perinatale uitkomsten en tweelingen- en familiegegevens.

## Overzicht van de thesis

In Hoofdstuk 2 focussen we specifiek op de analyse van tweelingen data. Deze spelen een belangrijke rol in het medisch onderzoek omdat ze een unieke bron van informatie bieden omtrent de impakt van genetische en omgevingsfactoren op het welzijn van de mens. We introduceren en onderzoeken eerst de structurele schattingsvergelijkingen (SEM), een methode die vaak gebruikt wordt voor het analyseren van tweelingen data en voor het schatten van de impakt van genetische factoren, i.e. erfelijkheid, op een bepaald fenotype. Het schatten van erfelijkheid brengt de vraag met zich mee of deze schatting vertekend kan zijn door de aanwezigheid van confounders. Dit werd, bij ons weten, nog slechts vaag beschreven in de literatuur. We onderzoeken dit daarom op het einde van Hoofdstuk 2 en geven een eenvoudige nieuwe methode om de schatting van erfelijkheid te beschermen tegen gemeten confounders.

In Hoofdstuk 3 onderzoeken we vervolgens, voor algemene geclusterde datastructuren, hoe schatters beschermd kunnen worden tegen de aanwezigheid van bepaalde confounders. We tonen aan hoe een statistische analyse die rekening houdt met de correlatie binnen clusters kan corrigeren voor ongemeten confounders die constant zijn binnen clusters. Tweelingen bijvoorbeeld delen logischerwijze dezelfde tweeling-specifieke kenmerken zodat een paarsgewijze vergelijking van tweelingen met onderling een verschillende blootstelling, niet verstoord wordt door de invloed van ongemeten tweeling-specifieke factoren. Dit begrip is welgekend in de statistiek en epidemiologie, maar niettemin maken heel wat frequent gehanteerde analyses daar geen gebruik van (bvb., veralgemeende schattingsvergelijkingen met onafhankelijke correlatie structuur). Ze bekomen op die manier onnodig

vertekende resultaten. Omdat correctie voor confounders zo belangrijk is voor een causale analyse gaan we na hoe dit gerealiseerd kan worden en onder welke voorwaarden dit opportuun is. Op die manier worden nieuwe schatters gecreëerd die onvertekend zijn zelfs wanneer er ongemeten confounders zijn die constant zijn binnen clusters. De bekomen nieuwe schattingsmethode (die we 'conditionele veralgemeende schattingsvergelijkingen' noemen) worden vervolgens aangewend om de validiteit na te gaan van een eenvoudige regressie methode waarin men een opsplitsing maakt tussen effecten binnen clusters en effecten tussen clusters (Neuhaus en Kalbfleisch, 1998). We tonen onder andere theoretisch aan dat de methode van Neuhaus en Kalbfleisch slechts consistente schatters levert bij lineaire modellen, en dus niet algemeen bruikbaar is. Door middel van simulaties tenslotte, vergelijken we beide methodes wat betreft vertekening en precisie van de schatters. We concluderen dat voor lineaire modellen, beide methodes vergelijkbaar zijn wat vertekening en precisie betreft, maar dat bij log-lineaire modellen bij de methode van Neuhaus en Kalbfleisch de schatters vertekend en minder precisie zijn.

Wegens de dringende onderzoeksvraag omtrent het effect van vruchtbaarheidsbehandelingen op perinatale uitkomsten vanuit de vakgroep Uro-Gynaecologie, gaan we vanaf Hoofdstuk 4 dieper in op het schatten van directe causale effecten. Dit zijn causale effecten van de blootstelling op de uitkomst die niet via gegeven andere variabelen (i.e. intermediaire variabelen) werken. Vruchtbaarheidsbehandelingen hebben immers via verschillende causale paden een effect op perinatale uitkomsten, bvb. door de kans op een tweeling te doen stijgen voor wie perinatale uitkomsten vaak slechter zijn dan voor eenlingen, door de grotere kans op een dizygote tweeling na deze behandeling, voor wie, vergeleken met monozygote tweelingen, de perinatale uitkomsten beter zijn, en daarnaast ook door een direct effect op deze uitkomsten.

In Hoofdstuk 4 kaderen we de complexiteit van het schatten van directe effecten (Pearl, 2001; Robins en Greenland, 1992; van der Laan en Petersen, 2004). Deze complexiteit ontstaat doordat het bepalen van het directe effect van een blootstelling op een uitkomst vereist dat men in de analyse corrigeert voor de gegeven intermediaire variabelen. Indien dat

via traditionele regressiemethoden gebeurt, ontstaan daardoor niet-causale associaties wanneer (zoals gewoonlijk) de intermediaire variabelen worden beïnvloed door prognostische factoren voor de uitkomst. Bijgevolg kunnen directe effecten doorgaans onmogelijk worden geschat via standaard statistische methoden, tenzij onder erg restrictieve assumpties (i.e. geen ongemeten confounders), die zelden realistisch zijn. Niettegenstaande deze restrictieve assumpties, wordt toch meestal traditionele regressie gebruikt voor het schatten van een direct effect en is de bekomen schatting in vele gevallen vertekend. Ten slotte tonen we in dit hoofdstuk dat het definiëren van een direct effect op zich al erg subtiel is en geven we enkele bestaande definities, waaronder het gecontrolleerde direct effect. Het is voor het schatten van dit gecontrolleerd direct effect dat we in de laatste 2 hoofdstukken nieuwe schattingsmethodes ontwikkelen

In Hoofdstuk 5 ontwikkelen we een eenvoudige schatter voor gecontrolleerde directe effecten die consistent is onder zwakkere assumpties dan de traditionele regressie-gebaseerde methodes voor directe effecten. Ze laten bijvoorbeeld toe dat confounders voor het effect van intermediaire variabelen op de uitkomst zelf beïnvloed zijn door de blootstelling. Door middel van simulaties en toepassing op een familie-gebaseerde genetische associatie studie, illustreren we de dramatische verbeteringen die bekomen worden door gebruik van deze schatter vergeleken met regressie-schatters. Bovendien is deze schatter zeer eenvoudig implementeerbaar in standaard statistische software.

In Hoofdstuk 6 tenslotte, tonen we dat de schatter uit Hoofdstuk 5 een bijzonder geval is van een meer algemene klasse van schatters voor directe effecten die gebaseerd zijn op invers wegen volgens een conditionele dichtheid van de intermediaire variabele. Wanneer deze variabele continu is, kunnen de gewichten en daarom dus ook de bekomen schattingen, erg onstabiel zijn. Om meer stabiele schattingen te bekomen, wordt daarom in dit hoofdstuk een dubbel robuuste schatter ontwikkeld. Deze is asymptotisch onvertekend als ofwel een conditioneel model voor de gemiddelde uitkomst correct gespecifieerd is of het model voor een conditionele dichtheid van de intermediaire variabele (dit zijn de gewichten). De schatter uit Hoofdstuk 5 wordt bekomen door het model voor de conditionele dichtheid van

de intermediaire variabele bewust verkeerd te specifiëren, namelijk door alle gewichten gelijk aan 1 te onderstellen, wat de schatting veel stabieler maakt. Deze schatting is consistent zolang het conditioneel model voor de gemiddelde uitkomst correct gespecifieerd is. Tenslotte gaan we nog een stap verder en ontwikkelen we andere dubbel robuuste schatters die zich zelfs in de aanwezigheid van extreme gewichten goed gedragen. De verschillende schatters worden vergeleken in uitvoerige simulatie studies en in de analyse van perinatale uitkomsten van eenlingen geboren na enkele of dubbele embryo transfer.

We besluiten de thesis met een overzicht en met een beschrijving van open onderzoeksproblemen die nauw gelinkt zijn aan de behandelde onderzoeksvragen.