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

Statistical methods for twin and sibling designs

Johan Zetterqvist

Stockholm 2017

Published by Karolinska Institutet. Printed by Eprint AB 2017

ISBN 978-91-7676-670-5

Karolinska Institutet

Institutionen för medicinsk epidemiologi och biostatistik

Statistical methods for twin and sibling designs

AKADEMISK AVHANDLING
som för avläggande av medicine doktorsexamen vid Karolinska Institutet offentligen försvaras i Atrium, Nobels väg 12B, Solna, Karolinska Institutet

Torsdagen den 11 maj, 2017, kl 09.00

av
Johan Zetterqvist
MSc

Huvudhandledare:
Docent Arvid Sjölander
Karolinska Institutet
Institutionen för medicinsk epidemiologi och biostatistik

Bihandledare:
Professor Yudi Pawitan
Karolinska Institutet
Institutionen för medicinsk epidemiologi och biostatistik

Professor Henrik Larsson
Örebro Universitet
Institutionen för medicinska vetenskaper samt
Karolinska Institutet
Institutionen för medicinsk epidemiologi och biostatistik

Professor Paul Lichtenstein
Karolinska Institutet
Institutionen för medicinsk epidemiologi och biostatistik

Fakultetsopponent:
Docent Theis Lange
Københavns Universitet
Institut for Folkesundhedsvidenskab
Biostatistisk afdeling

Betygsnämnd:
Docent Ingeborg Waernbaum
Umeå universitet
Handelshögskolan vid Umeå universitet
Enheten för Statistik

Docent Michael Höhle
Stockholms Universitet
Matematiska Institutionen
Avdelningen för matematisk statistik

Docent Fredrik Granath
Karolinska Institutet
Institutionen för medicin, Solna
Enheten för Klinisk epidemiologi

Stockholm 2017

Abstract

Twin and sibling studies are valuable in that they allow adjustment for potential confounding factors that are impossible or hard to measure. By measuring associations 'within-cluster' it is possible to adjust for many factors that are shared between individuals in the same cluster.

Using Swedish national registers, it is possible to obtain information about a large number of potential confounders. While this gives medical researchers great opportunities to control for confounding, it also increases the risk of model misspecification leading to biased estimates. One strategy to reduce the risk of such bias is to use doubly robust (DR) estimation. In DR estimation two working models are combined in such a way that the resulting estimate will remain asymptotically unbiased when one of the models is misspecified.

In **study I**, we implement existing DR estimators for parameters in linear, log-linear and logistic regression models in the `R` package `drgee`. In **study II**, we propose a new class of DR estimators for 'within-cluster' association measures in linear and log-linear regression models. In **study III** we propose a DR estimator for the 'within-cluster' log odds ratio parameter in logistic regression models. The estimators proposed in studies II and III are also implemented in the `R` package `drgee`.

In **study IV**, we discuss what shared factors the 'within-cluster' association actually is adjusted for. Using the formal theory of causal diagrams we demonstrate that the standard methods for estimating 'within-cluster' association parameters implicitly adjust for shared confounders, shared mediators, but not shared colliders. Therefore, the estimated parameter may have a causal interpretation as a direct effect, i.e. as the part of the causal effect that is not mediated through shared factors.

List of scientific papers

The following papers, referred to in the text by their Roman numerals, are included in this thesis.

I Johan Zetterqvist and Arvid Sjölander.
Doubly robust estimation with the R package drgee.
Epidemiologic Methods, 4(1):69–86, 2015.

II Johan Zetterqvist, Stijn Vansteelandt, Yudi Pawitan, and Arvid Sjölander.
Doubly robust methods for handling confounding by cluster.
Biostatistics, 17(2):264–276, 2016.

III Johan Zetterqvist, Karel Vermeulen, Stijn Vansteelandt, and Arvid Sjölander.
Doubly robust conditional logistic regression.
2017. Submitted.

IV Arvid Sjölander and Johan Zetterqvist.
Confounding, mediation and colliding in conditional maximum likelihood estimation.
Epidemiology, 2016. Accepted.

Contents

1 Introduction **1**

2 Background **2**

2.1 Causation versus statistical association in observational studies 2

2.2 Causal models and effect parameters 3

2.2.1 Assumptions 3

2.2.2 Causal and associational models 4

2.2.3 Different effect measures 4

2.2.4 Effect modification 5

2.2.5 Directed acyclic graphs 5

2.3 Dealing with measured confounders 6

2.3.1 Generalized linear models 6

2.3.2 M-estimation 6

2.3.3 Target parameters vs nuisance parameters 7

2.3.4 G-estimation 8

2.3.5 G-estimation versus inverse probability weighting 9

2.3.6 Non-collapsibility 10

2.4 Doubly robust estimation 11

2.4.1 Bias due to model misspecification 11

2.4.2 Doubly robust estimators 11

2.4.3 The doubly robust G-estimator 12

2.4.4 Doubly robust logistic regression 12

2.5 Methods for clustered data 13

2.5.1 Clustered data 13

2.5.2 Maximum likelihood estimation using clustered data 14

2.5.3 Conditional maximum likelihood 15

2.5.4 Conditional generalized estimating equations 16

2.5.5 Underlying assumptions 16

3 Aims **18**

4 Summary of papers **19**

4.1 Paper I 19

4.1.1 Doubly robust G-estimation and G-estimation 19

4.1.2 Doubly robust estimation for logistic models 20

4.1.3 The `R` package `drgee` 20

4.2 Paper II 21

4.3 Paper III 22

4.4 Paper IV 23

5 Discussion **26**

5.1 Limitations 28

5.2 Extensions 28

6 Acknowledgements **31**

List of abbreviations

RCT	randomized controlled trial
ADHD	attention deficit-hyperactive disorder
DAG	directed acyclic graph
GLM	generalized linear model
ML	maximum likelihood
GEE	generalized estimating equations
DR	doubly robust
CML	conditional maximum likelihood
CGEE	conditional generalized estimating equations

1 Introduction

One of the main goals of epidemiology and of medical research is to find out how to improve public health by removing or reducing factors that have a negative impact on public health and increase factors that are beneficial to public health. Even though randomized clinical trials (RCTs) are often the preferred way to identify such factors, RCTs are often not feasible (economically, practically or ethically). Further, since participants in RCTs are not a random sample, the results from RCTs can not be generalized to the target population. Therefore, one often have to use observational data to find such factors. Doing so can be a challenging task, since the associations of interest might be 'distorted' by other factors. For instance, in Sweden there is a significant positive association between prescription of stimulant medication and criminality. This does not mean that we can reduce the rate of crime by stopping all prescriptions of stimulant medication. Stimulant medication is mainly prescribed for individuals that are diagnosed with attention deficit-hyperactive disorder (ADHD). At the same time, ADHD is a risk factor for criminal behavior. Thus the observed crude association may partly be explained by ADHD being a common cause for stimulant medication and criminal behavior. When the aim is to study the causal effect of stimulant medication on criminal behavior, the crude association is misleading. Such associations are said to be 'confounded' by common causes (see e.g. Pearl, 2009).

Confounded associations are common in observational data. In order to draw causal conclusions from observational data, it is therefore necessary to 'adjust' for potential confounding factors, e.g. by performing analyses stratified on such factors or by using the factors as covariates in a regression model. For instance, after adjustment for individual-specific factors (e.g. ADHD diagnose) the association between stimulant medication and criminal behavior becomes reversed (Lichtenstein et al., 2012).

In this thesis, new statistical methods are presented. These methods are intended to facilitate adjustment for confounding in obsevational studies; in particular in studies using clustered data, a large number of measured covariates, and a large number of observations.

2 Background

2.1 Causation versus statistical association in observational studies

Although the statistical methods in this thesis can be described without referring to questions about causality, the methods are motivated by questions about causality. Suppose that we are interested in whether an exposure X has a causal effect on an outcome Y. Given observed data, what can be said about this effect? A first approach could be to compare the mean values of Y under two different levels of X, say

$$g\{\mathrm{E}(Y|X=x)\} - g\{\mathrm{E}(Y|X=0)\} = \beta^* x \ , \tag{1}$$

where g is some link function. Then β^* is a measure of the statistical association between X and Y. For instance, assume that both Y and X are binary and that g the identity link. Then β^* quantifies the difference in risk of having the outcome comparing exposed individuals to unexposed individuals. When g is the log link, β^* quantifies the corresponding log risk ratio instead. In an RCT with perfect compliance and no dropouts, a statistical test of whether $\beta^* = 0$ is equivalent to a test of whether X has a causal effect on Y. In an observational study, this may no longer be true. For instance, there may be a set of variables $\mathbf{V} = (V_1, \ldots, V_p)$ that causes both X and Y. In such a scenario, we might detect a statistical association between X and Y even if there is no causal effect of X on Y. In that case we say that the studied association is 'confounded' and we refer to the elements of $\mathbf{V}$ as the 'confounders'. In such a scenario, the parameter β^* in the associational model (1) is not the parameter we are interested in. In contrast, a 'causal model' is a model for the causal effect. It can be formulated as

$$g\{\mathrm{E}(Y^x)\} - g\{\mathrm{E}(Y^0)\} = \beta_c^* x, \tag{2}$$

where $\mathrm{E}(Y^x)$ is the mean outcome that we would observe if all subjects were assigned the exposure $X = x$. We refer to β_c^* as the 'causal effect parameter' and to the effect it measures as the 'causal effect'.

2.2 Causal models and effect parameters

2.2.1 Assumptions

In order to interpret the parameter β^* in (1) as the causal effect parameter β_c^*, we need two assumptions to hold - 'consistency' and 'exchangeability'. 'Consistency' means that the observed Y for subjects with exposure level x would have the same value regardless of whether the exposure level was just observed to be x or whether it was forced to the level x, i.e.

$$X = x \qquad \Longrightarrow \qquad Y = Y^x \ . \tag{3}$$

This assumption would be violated if the (counter-factual) assignment mechanism would have an effect on the outcome that does not go through the treatment X. Henceforth we will assume that (3) always holds.

'Exchangeability' means that there are no systematic differences between subjects with different levels of exposure. Another way to formulate this is that the observed exposure level is independent of any outcome that would have been observed if the exposure level was forced to some level, i.e.

$$Y^x \perp\!\!\!\perp X \text{ for all } x. \tag{4}$$

If there are common causes $\mathbf{V}$ of X and Y, i.e. if the association (1) is confounded by $\mathbf{V}$, this assumption is generally violated. This is so because the distribution of Y^x depends on the distribution of $\mathbf{V}$, which in turn is associated with the observed distribution of X. For instance if the exposure X is stimulant medication, Y is criminal behavior, and V is the severity of the disorder, then V would still be associated with the outcome in the counter-factual scenario where all subjects were medicated while also being associated with the observed exposure X. However, if there are no other common causes of X and Y,

$$Y^x \perp\!\!\!\perp X|\mathbf{V} \text{ for all } x \ . \tag{5}$$

In other words, for a group of individuals with the same level of $\mathbf{V}$, there is no association between the observed exposure X and the outcome Y^x that would be observed

if everyone was forced to the exposure level x. We call this 'conditional exchangeability'. If conditional exchangeability holds, we say that the set $\mathbf{V}$ is sufficient to control for confounding. To simplify the presentation, we will also call the members of the set $\mathbf{V}$ 'confounders', thus we are implicitly assuming that this set is the only set that is sufficient to control for confounding. For a more stringent definition of confounding and confounders, see Pearl (2009).

2.2.2 Causal and associational models

If conditional exchangeability (5) holds, we can formulate a *causal* model

$$g\{\mathrm{E}(Y^x|\mathbf{V})\} - g\{\mathrm{E}(Y^0|\mathbf{V})\} = \beta_c x \ . \tag{6}$$

where β_c has a causal interpretation as the difference on the scale of the link function g, for each value of $\mathbf{V}$, comparing the expected outcome if every individual was assigned the treatment level x with the expected outcome if every individual as assigned the treatment level 0. If consistency (3) also holds, the associational model

$$g\{\mathrm{E}(Y|X = x, \mathbf{V})\} - g\{\mathrm{E}(Y|X = 0, \mathbf{V})\} = \beta x \tag{7}$$

can be used to obtain an estimate of the causal parameter β_c, since

$$\begin{aligned}\beta_c x &\overset{(6)}{=} g\{\mathrm{E}(Y^x|\mathbf{V})\} - g\{\mathrm{E}(Y^0|\mathbf{V})\} \\ &\overset{(5)}{=} g\{\mathrm{E}(Y^x|X = x, \mathbf{V})\} - g\{\mathrm{E}(Y^0|X = 0, \mathbf{V})\} \\ &\overset{(3)}{=} g\{\mathrm{E}(Y|X = x, \mathbf{V})\} - g\{\mathrm{E}(Y|X = 0, \mathbf{V})\} \overset{(7)}{=} \beta x \ ,\end{aligned}$$

where the numbers above the equality signs refer to the causal model assumption (6), the assumption of conditional exchangeability (5), the assumption of consistency (3), and the associational model assumption (7), respectively. We will refer to β_c as an 'effect measure'.

2.2.3 Different effect measures

Since β_c measures the causal effect on the scale of the link function g, the interpretation of β_c also depends on g. Suppose that both exposure X and outcome Y are binary. Then, if g is the identity link β_c can be interpreted as a causal risk difference.

In contrast, if g is the log link, e^{β_c} can be interpreted as a causal risk ratio. When g is the logit link, e^{β_c} has an interpretation as a causal odds ratio.

Sometimes one is interested in the causal *marginal* effect parameter $\beta_c^* = g\{\mathrm{E}(Y^{x+1})\} - g\{\mathrm{E}(Y^x)\}$, measuring the effect of increasing the exposure level with one unit, rather than the *conditional* effect β_c in (6). When g is the identity link, this estimand is denoted the 'average treatment effect'. Often, the adjusted marginal effect parameter β_c^* can be obtained from the conditional effect parameter β_c in (6), by standardization over $\mathbf{V}$ (Robins, 1986). However, in this thesis we will only be concerned with the conditional effect parameter, partly because we will include scenarios in which standardization is not feasible.

2.2.4 Effect modification

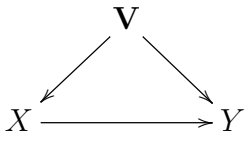
In the model (6), we assume that the association is the same for all levels $\mathbf{v}$ of $\mathbf{V}$ such that $\Pr(\mathbf{V} = \mathbf{v}) > 0$ (or $f_{\mathbf{V}}(\mathbf{v}) > 0$). To relax this assumption, we can formulate the causal model

$$g\{\mathrm{E}(Y^x|\mathbf{V})\} - g\{\mathrm{E}(Y^0|\mathbf{V})\} = \boldsymbol{\beta}_c\mathbf{Z}x \ , \tag{8}$$

where $\mathbf{Z} = \mathbf{Z}(\cdot)$ is some vector valued function of $\mathbf{V}$. In order to parametrize the main effect of x, one can let $\mathbf{Z}$ have 1 as first element. To improve readability of the presentation, we will henceforth only consider the case when β_c (and β) remain the same for all levels of $\mathbf{V}$. However, all models and estimators considered in this thesis can be extended to allow for effect modification by $\mathbf{V}$ by replacing β with $\boldsymbol{\beta}\mathbf{Z}(\mathbf{V})$.

2.2.5 Directed acyclic graphs

A common way to visualize causal relations is to use a directed acyclic graph (DAG) (Pearl, 1995). A causal effect of X on Y confounded by $\mathbf{V}$ can thus be pictured as



.

In a DAG, the *presence* of a directed edge (arrow) symbolizes a *possible* causal effect. The *absence* of a directed edge symbolizes an *absence* of a causal effect. For instance

the presence of a directed edge from $\mathbf{V}$ to X indicates that there is a possible causal effect of $\mathbf{V}$ on X.

In this thesis, we will put parameters next to edges in DAGs as representations of quantified associations. The parameters themselves will be more precisely defined in the text. We hope that this slight abuse of DAGs will enhance understanding, rather than the opposite.

2.3 Dealing with measured confounders

2.3.1 Generalized linear models

For point outcomes, one of the most common classes of models in the epidemiology is the generalized linear model (GLM), i.e.

$$Y|X, \mathbf{V} \sim P_{Y|X,\mathbf{V};\boldsymbol{\theta},\boldsymbol{\omega}} \tag{9a}$$

$$g\{\mathrm{E}(Y|X, \mathbf{V})\} = \beta X + \mu + \boldsymbol{\gamma}\mathbf{V} \ , \tag{9b}$$

where Y is the outcome interest, X the exposure, $\mathbf{V}$ is a vector of covariates, g is some specified (continuous and invertible) link function, $P_{Y|X,\mathbf{V};\boldsymbol{\theta},\boldsymbol{\omega}}$ is some specified probability distribution in the exponential family, indexed by the parameters $\boldsymbol{\theta} = (\beta, \mu, \boldsymbol{\gamma})$ and $\boldsymbol{\omega}$. We will refer to the second part - (9b) - as a 'regression model'. The maximum likelihood (ML) estimate $\hat{\boldsymbol{\theta}}_{ML}$ is obtained by solving the score equations

$$\sum_{i=1}^{n} \frac{\partial}{\partial \boldsymbol{\theta}} \log p_{Y|X,\mathbf{V};\boldsymbol{\theta},\boldsymbol{\omega}}(y_i|x_i, \mathbf{v}_i) = 0 \tag{10}$$

for $\boldsymbol{\theta}$. When $P_{Y|\mathbf{V};\boldsymbol{\theta},\boldsymbol{\omega}}$ is the normal distribution with homoscedastic errors and $g(\cdot)$ is the identity link, the score equations (10) have the form

$$\sum_{i=1}^{n} \begin{pmatrix} x_i \\ 1 \\ \mathbf{v}_i \end{pmatrix} \{y_i - (\beta x_i + \mu + \boldsymbol{\gamma}\mathbf{v}_i)\} = 0 \ . \tag{11}$$

2.3.2 M-estimation

Under some mild regularity conditions (see e.g. van der Vaart, 2000), the solution $\hat{\boldsymbol{\theta}}_{ML}$ to (11) is consistent for the true value of $\boldsymbol{\theta}$ in the model (9b) regardless of whether

the distributional assumption (9a) is correct, as long as the regression model (9b) is correct. Similarly, for any continuous and invertible link function $g(\cdot)$, solving the estimating equations

$$\sum_{i=1}^{n} \begin{pmatrix} x_i \\ 1 \\ \mathbf{v}_i \end{pmatrix} \left\{ y_i - g^{-1}(\beta x_i + \mu + \boldsymbol{\gamma}\mathbf{v}_i) \right\} = 0 \tag{12}$$

for $\boldsymbol{\theta}$ gives a consistent estimate $\hat{\boldsymbol{\theta}}$ of the true value of $\boldsymbol{\theta}$ when the regression model (9b) is correct, even when the distributional assumption (9a) is incorrect. Estimating equations of the form (12) are called generalized estimating equations (GEEs). Often, GEEs are written in a more general form that also allows data to be clustered (Liang and Zeger, 1986).

Estimators based on estimating equations (12) are examples of so called M-estimators (Stefanski and Boos, 2002). M-estimators can be thought of as generalizations of ML estimators, where the score functions $\frac{\partial}{\partial \boldsymbol{\theta}} \log p_{Y|X,\mathbf{V};\boldsymbol{\theta},\boldsymbol{\omega}}(y_i|x_i, \mathbf{v}_i)$ are replaced by functions $M(y_i|x_i, \mathbf{v}_i; \boldsymbol{\theta}, \boldsymbol{\omega})$ with the property that the equation $\mathrm{E}\left\{M(Y_i|\mathbf{x}_i; \boldsymbol{\theta}, \boldsymbol{\omega})\right\} = 0$ have unique solution $\boldsymbol{\theta} = \boldsymbol{\theta}_0$. The parameter $\boldsymbol{\theta}$ is estimated by solving the estimating equations

$$\sum_{i=1}^{n} M(y_i|x_i, \mathbf{v}_i; \boldsymbol{\theta}, \boldsymbol{\omega}) = 0$$

for $\boldsymbol{\theta}$. Assuming regularity conditions (see e.g. van der Vaart, 2000), the distribution of the resulting estimate $\hat{\boldsymbol{\theta}}_n$ can be shown to be asymptotically normal with a variance that can be obtained using a sandwich estimator (Stefanski and Boos, 2002).

2.3.3 Target parameters vs nuisance parameters

In real scenarios, not all parameters in a regression model are of interest. Often, one is only interested in the association between two variables. The corresponding parameter are therefore termed 'the target parameter'. Often we want give the target parameter a causal interpretation. Other independent variables may be used in order to adjust for confounding only. The corresponding parameters are then not of primary interest and are therefore called 'nuisance parameters'. Assuming that we

are interested in the association between an exposure X and an outcome Y, adjusted for a set of covariates $\mathbf{V}$, we can formulate a regression model

$$g\{\mathrm{E}(Y|X,\mathbf{V})\} = \beta X + \boldsymbol{\gamma}\mathbf{V} \ . \tag{13}$$

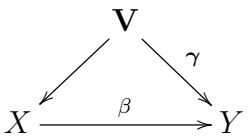
Then β is the target parameter and $\boldsymbol{\gamma}$ is the nuisance parameter. The model (13) can be split into two parts. The first part

$$g\{\mathrm{E}(Y|X,\mathbf{V})\} - g\{\mathrm{E}(Y|X=0,\mathbf{V})\} = \beta X \tag{14}$$

models the effect of X on Y, adjusted for $\mathbf{V}$ and contains the target parameter β. This model will be referred to as the 'main model'. The second part

$$g\{\mathrm{E}\{Y|X=0,\mathbf{V})\} = \boldsymbol{\gamma}\mathbf{V} \ , \tag{15}$$

models the distribution of Y conditional on $X = 0, \mathbf{V}$ and specifies how to adjust for $\mathbf{V}$. This model contains the nuisance parameter $\boldsymbol{\gamma}$ and will be referred to as the 'outcome nuisance model'. The two models can be visualized as the parameter symbols β and $\boldsymbol{\gamma}$ next to the $X \to Y$ and the $\mathbf{V} \to Y$ arrow, respectively, in the DAG



.

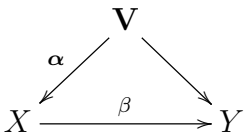
In order to consistently estimate β in (14) using the regression model (13), it is required that the outcome nuisance model (15) is correctly specified. Therefore, it may be motivated to allow the outcome nuisance model to have a more flexible formulation by replacing $\boldsymbol{\gamma}\mathbf{V}$ with a more general function $f_O(\mathbf{V};\boldsymbol{\gamma})$ of $\mathbf{V}_{ij}$. However, since it is common practice to only include the main effects in the regression model and to make notation easier, we will henceforth use (15) as our outcome nuisance model. However, all considered nuisance models can easily be extended to include more general functions.

2.3.4 G-estimation

An alternative strategy to estimate β in (14) is to augment the model (14) with a model for the exposure, e.g.

$$h\{\mathrm{E}(X|\mathbf{V})\} = \boldsymbol{\alpha}\mathbf{V} \ , \tag{16}$$

where h is some link function. Using G-estimation we can then obtain an estimate of β by solving estimating equations based on the main model (14) and the exposure nuisance model (16) (Robins et al., 1992). Since the model (16) is only used in order to adjust for $\mathbf{V}$, the parameter $\boldsymbol{\alpha}$ is considered to be a nuisance parameter and the model itself will be referred to as 'the exposure nuisance model'. The main model and the exposure nuisance model can be visualized as the $X \to Y$ and the $\mathbf{V} \to Y$ arrow, respectively, in the DAG



.

One limitation of G-estimation is that for other link functions $g(\cdot)$ than identity and log, further distributional assumptions are required (Vansteelandt et al., 2014). G-estimation may be preferable when the researcher have better information about how the exposure depends on the covariates $\mathbf{V}$. However, when the exposure nuisance model is misspecified, G-estimation can give biased estimates of β. As with the outcome nuisance model (15), more general functions $f_E(\mathbf{V}; \boldsymbol{\alpha})$ of $\mathbf{V}$ in (16) may be motivated in practice, but we will use this simple form to keep notation simple.

2.3.5 G-estimation versus inverse probability weighting

Adjustment for potential confounding by modeling the exposure as a function of potential confounding factors can also be done by using with inverse probability weighting (IPW). While both G-estimation and IPW-based estimation can be used to estimate adjusted treatment effects, they do not target the same parameter. While IPW-based estimation targets the marginal causal effect β_c^*, G-estimation targets the causal effect conditional on confounding factors β_c. Thus the two methods are used for different research questions. Still, IPW-based methods are far more popular than methods based on G-estimation, even though G-estimation have several advantages (Vansteelandt et al., 2014; Robins, 2000a). A partial explanation for this is that, until recently, there has been a lack of software implementations for G-estimation.

2.3.6 Non-collapsibility

An important concept to distinguish from confounding is non-collapsibility. According to one definition of collapsibility, a regression model is (strictly) collapsible for a parameter β over a set of covariates $\mathbf{V}$ if β remains the same in the corresponding regression model with $\mathbf{V}$ omitted (Greenland et al., 1999; Pearl, 2009). For instance, the model (13) is collapsible for β over $\mathbf{V}$ if β is the same as β^* in the marginal (over $\mathbf{V}$) regression model

$$g\{\mathrm{E}(Y|X=x)\} = \nu^* + \beta^* x \ . \tag{17}$$

Sometimes non-collapsibility of (13) for β over $\mathbf{V}$ is interpreted as confounding. However, confounding might not be the only reason that $\beta \neq \beta^*$. As noted by Janes et al. (2010), the difference between β and β^* can be separated into two parts. By defining β as the adjusted parameter in the conditional main model (14), β^* as the parameter in the marginal (crude) model (17), β_c^* as the marginal causal parameter defined in the marginal causal model (2), and β_c as the conditional causal parameter defined in the conditional causal model (6), we can write

$$\beta^* - \beta = (\beta^* - \beta_c^*) + (\beta_c^* - \beta) \ .$$

The difference $\beta^* - \beta_c^*$ can be interpreted as the part that is due to confounding, since it is the difference between a crude parameter and a causal parameter. Under conditional exchangeability, (5), β is the same as the conditional causal parameter β_c. Therefore, the second difference $\beta_c^* - \beta$ can be interpreted as a difference $\beta_c^* - \beta_c$ between two causal parameters. This difference quantifies 'the non-linearity effect' and is often related to the type of link function g used. For instance, when g is the identity or log link function, $\beta_c^* = \beta_c$. In contrast, when g is the logit link, $|\beta_c^*| \geq |\beta_c|$ (with equality only if $\mathrm{Var}\,\mathbf{V} = 0$ or if $\boldsymbol{\gamma} = 0$) (Neuhaus and Jewell, 1993). Thus, adjustment for a covariate can also alter the association measure itself. For this reason, the odds ratio is sometimes referred to as 'non-linear' (Janes et al., 2010) or 'non-collapsible' (Hernán et al., 2011; Greenland and Pearl, 2011; Pearl, 2009).

2.4 Doubly robust estimation

2.4.1 Bias due to model misspecification

When a model is used to estimate a parameter, the assumed model needs to be correct. In other words, the model needs to be a correct description of the data. For instance, assume that Y, conditional on X and $\mathbf{V}$, is normally distributed with mean

$$\mathrm{E}(Y|X=x,\mathbf{V}=\mathbf{v}) = \beta x + e^{\nu+\delta\mathbf{v}} \ . \tag{18}$$

If we (incorrectly) assume that

$$\mathrm{E}(Y|X=x,\mathbf{V}=\mathbf{v}) = \mu + \beta x + \boldsymbol{\gamma}\mathbf{v} \tag{19}$$

the ML estimator $\hat{\beta}_{ML}$ will not be consistent for the parameter β in the main model

$$\mathrm{E}(Y|X=x,\mathbf{V}=\mathbf{v}) - \mathrm{E}(Y|X=0,\mathbf{V}=\mathbf{v}) = \beta x \ . \tag{20}$$

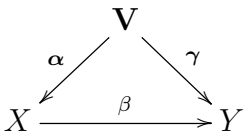
This is because the misspecification of the nuisance model $\mathrm{E}(Y|X=0,\mathbf{V}=\mathbf{v})$ 'spills over' to the estimate of β, causing inconsistent estimates of β. In practice, most of the models are, to some extent, incorrectly specified. Therefore, this kind of bias is always present. There are several strategies to reduce bias due to model specification. One option is to estimate the nuisance models non-parametrically or by using splines. Although such estimators exist, they tend to perform badly when $\mathbf{V}$ is high-dimensional, even in moderate sample sizes (Robins and Ritov, 1997).

2.4.2 Doubly robust estimators

A middle way between using a non-parametric nuisance model and relying on a nuisance model, is to use a doubly robust (DR) estimator. The notion was introduced Scharfstein et al. (1999) to denote estimators that combines two working nuisance models such that a consistent estimate of the target parameter is obtained when at least one of the two nuisance models is correct, not necessarily both. The detailed mathematical theory underlying the methodology has been described by Robins and Rotnitzky (2001) and van der Laan and Robins (2003). A more general overview of the methodology DR estimation was given by Bang and Robins (2005).

2.4.3 The doubly robust G-estimator

The G-estimator, described in Section 2.3.4, can be extended to also incorporate an outcome nuisance model of the form (15). The resulting estimator, proposed by Robins (2000b), has the property of being DR, i.e. it will remain consistent for β in the main model (14) if either the outcome nuisance model (15) or the exposure nuisance model (16) is correctly specified. The scenario can be visualized in the following figure, where the parameters $\boldsymbol{\alpha}$, $\boldsymbol{\gamma}$, and β next to the edges in the DAG symbolizes the three models (16), (15), and (14), respectively:



.

2.4.4 Doubly robust logistic regression

When g is the logit link, the regression model (13) have the form

$$\text{logit}\{\text{E}(Y|X,\mathbf{V})\} = \beta X + \boldsymbol{\gamma}\mathbf{V} \ .$$

When the outcome Y is binary, this model is more commonly formulated in terms of probabilities for the outcome

$$\text{logit}\{\Pr(Y=1|X,\mathbf{V})\} = \beta X + \boldsymbol{\gamma}\mathbf{V} \ . \tag{21}$$

Unfortunately, DR G-estimation is not useful for logistic regression models (Robins, 2000a). However, when both the outcome Y and the exposure X are binary, the log odds ratio β can also be modeled as

$$\text{logit}\{\Pr(X=1|Y,\mathbf{V})\} = \beta Y + \boldsymbol{\alpha}\mathbf{V} \ , \tag{22}$$

where $\mathbf{V}$ is some vector valued function of $\mathbf{V}$ (Prentice, 1976). The models (21) and (22) both share the main model

$$\frac{\Pr(Y=1,X=1|\mathbf{V})\Pr(Y=0,X=0|\mathbf{V})}{\Pr(Y=1,X=0|\mathbf{V})\Pr(Y=0,X=1|\mathbf{V})} = e^{\beta} \tag{23}$$

for the odds ratio conditional on $\mathbf{V}$. The models (21) and (22) have been termed 'the prospective logistic model' and 'the retrospective logistic model' respectively (Breslow

and Powers, 1978). Even though prospective and retrospective logistic models contain the same log odds parameter β, they use different nuisance models to adjust for $\mathbf{V}$. The prospective logistic model uses the model

$$\text{logit}\{\Pr(Y = 1|X = 0, \mathbf{V})\} = \boldsymbol{\gamma}\mathbf{V} \tag{24}$$

for the influence of $\mathbf{V}$ on the proportion affected among the unexposed. The retrospective logistic model uses the model

$$\text{logit}\{\Pr(X = 1|Y = 0, \mathbf{V})\} = \boldsymbol{\alpha}\mathbf{V} \tag{25}$$

for the influence of $\mathbf{V}$ on the proportion exposed among those unaffected by the outcome. In sampling designs such as matched case control studies and matched cohort studies, the distributions modeled by (24) and (25) are clearly not representative of the target population and are therefore not of interest. In contrast, the odds ratio in the sample is representative of the odds ratio in the target population in such sampling designs. Nevertheless, misspecification of the nuisance model when doing prospective or retrospective logistic regression may lead to biased estimates of β.

By utilizing the symmetry of the conditional odds ratio one can construct a DR estimator of β , i.e. it is consistent for β in (37) if either the outcome nuisance model (24) and or the exposure nuisance model (25) is correctly specified, not necessarily both (Chen, 2007). This estimator can be formulated as a two-step estimator (Tchetgen Tchetgen et al., 2010), where estimates $\hat{\boldsymbol{\gamma}}$ and $\hat{\boldsymbol{\alpha}}$ of the nuisance parameters are obtained in the first step, e.g. using ML estimation. In the second step, a DR estimate of the log odds ratio β in the main model (37) is obtained using the estimates $\hat{\boldsymbol{\gamma}}$ and $\hat{\boldsymbol{\alpha}}$ from the first step.

2.5 Methods for clustered data

2.5.1 Clustered data

Most statistical methods are concerned with independent data. When dealing with clustered data, special methods are required to deal with the dependence between observations. For instance, when using (GEE) to estimate parameters from clustered

data, the estimate of the variance of the estimated coefficients needs to be adjusted for the clustering (Liang and Zeger, 1986). However, the clustering may not only be a nuisance. The information about clustering can also be of advantage since it allows us to adjust for unmeasured factors that are shared between members of the same cluster. Henceforth, we assume that data can be structured into M distinct units indexed by i, each consisting of n_i observations. For each observation j in cluster i, we assume that we can observe an outcome Y_{ij} , an exposure X_{ij}, and a set of covariates $\mathbf{V}_{ij}$. In addition, we also assume a set of factors $\mathbf{W}_i$ that are shared between members of the same cluster i and may be partially unobserved. We will henceforth assume that observations from different clusters are independent.

In matched case-control and matched cohort studies, $\mathbf{W}_i$ consists of matching variables. Since matching variables are measured, they can be explicitly adjusted for in a regression model, e.g.:

$$g\{\mathrm{E}(Y_{ij}|\mathbf{X}_i, \mathbf{V}_i, \mathbf{W}_i)\} = \mu + \beta X_{ij} + \boldsymbol{\gamma}\mathbf{V}_{ij} + \boldsymbol{\delta}\mathbf{W}_i \tag{26}$$

for $i = 1, \ldots, M$ and $j = 1, \ldots, n_i$ where n_i is the size of cluster i, g is some link function and where $\mathbf{X}_i$ and $\mathbf{V}_i$ are all the exposures and covariates, respectively, in cluster i. In other studies with clustered data, $\mathbf{W}_i$ may be only partially observed; e.g. in co-twin control studies, in sibling studies, and in studies with repeated measures. In those cases it is common to assume a more general model

$$g\{\mathrm{E}(Y_{ij}|\mathbf{X}_i, \mathbf{V}_i, \mathbf{W}_i)\} = \mu_i + \beta X_{ij} + \boldsymbol{\gamma}\mathbf{V}_{ij} \tag{27}$$

where the effect of $\mathbf{W}_i$ is absorbed into the cluster-specific intercept μ_i. The model (26) above is a special case of this model with $\mu_i = \mu + \boldsymbol{\delta}\mathbf{W}_i$.

2.5.2 Maximum likelihood estimation using clustered data

Using observations from M clusters, the parameters $\beta, \boldsymbol{\gamma}, \mu_1, \ldots, \mu_M$ can be estimated with ML. However, the ML estimate $\hat{\beta}_{ML}$ of β in the model (27) is not consistent in

general. The consistency of $\hat{\beta}_{ML}$ partly depends on the way additional data is sampled. If the cluster size n_i grows with the number of observations while the number of clusters M remain fixed there will be a fixed number of parameters $(\beta, \boldsymbol{\gamma}, \mu_1, \ldots, \mu_M)$ to estimate and the ML estimate $\hat{\beta}_{ML}$ will be consistent. This is the case in studies with a fixed cohort of M individuals and where new data is obtained by additional measurements of the same individuals. In contrast, if the cluster sizes n_i are expected to remain the same while the number of clusters M will increase, as in twin and sibling studies, the number of parameters will grow with the sample size and the ML estimate is no longer guaranteed to be consistent (Neyman and Scott, 1948; Breslow, 1981). In general, ML estimation is most appropriate with a small number of large clusters (counties in Sweden, municipalities). When there are a large number of small clusters (e.g. twins, siblings, school classes), other estimation methods may be needed.

2.5.3 Conditional maximum likelihood

For the case when the number of clusters grows with the sample size, conditional maximum likelihood (CML) is often used (Andersen, 1970). Assuming a distribution for Y_i and regression model

$$g\{\mathrm{E}(Y_{ij}|\mathbf{X}_i, \mathbf{V}_i, \mathbf{W}_i)\} = \mu_i + \beta X_{ij} + \boldsymbol{\gamma}\mathbf{V}_{ij} \tag{28}$$

it is possible to estimate β if there exists sufficient statistics $T_1, T_2, \ldots$ for the cluster-specific intercepts $\mu_1, \mu_2, \ldots$. The conditional likelihood

$$\prod_i p\left(\mathbf{Y}_i|\mathbf{X}_i, \mathbf{V}_i, T_i; \beta, \boldsymbol{\gamma}\right)$$

is then free of the cluster-specific intercepts and we can estimate β by maximizing this likelihood with respect to β and γ. For a GLM with canonical link g, one can use $T_i = \sum_{j=1}^{n_i} Y_{ij}$ (McCullagh and Nelder, 1989). Conditional maximum likelihood (CML) estimation is used to estimate the parameters in conditional logistic regression and in conditional Poisson regression.

2.5.4 Conditional generalized estimating equations

When g in (48) is the identity or log link, it is also possible to estimate β in (48) by solving 'conditional generalized estimating equations' (CGEEs) (Goetgeluk and Vansteelandt, 2008). By letting

$$S_{Y,ij}(\beta, \boldsymbol{\gamma}) = \begin{cases} Y_{ij} - \beta X_{ij} - \boldsymbol{\gamma}\mathbf{V}_{ij} & \text{when } g \text{ is the identity link} \\ Y_{ij} e^{-\beta X_{ij} - \boldsymbol{\gamma}\mathbf{V}_{ij}} & \text{when } g \text{ is the log link} \end{cases}$$

we can obtain consistent estimates of β and $\boldsymbol{\gamma}$ by solving estimating equations of the form

$$\sum_{i=1}^{M} \sum_{j=1}^{n_i} \left(d_{ij} - \bar{d}_i\right) S_{Y,ij}(\beta, \boldsymbol{\gamma}) = 0$$

for β, where d_{ij} is some function of X_{ij} and $\mathbf{V}_{ij}$ and where $\bar{d}_i = \frac{1}{n_i} \sum_{k=1}^{n_i} d_{ik}$. This kind of estimators require less strict assumptions than CML estimators. Further, it can be shown that CML estimators are asymptotically equivalent to subgroups (i.e. with appropriate choices of d_{ij}) of this kind of estimators. A drawback is that the they do not work when g is the logit link (Goetgeluk and Vansteelandt, 2008).

2.5.5 Underlying assumptions

Most statistical methods are designed with independent data in mind. When data are clustered, it is important to consider the assumptions that underlie the methods, since violations of the underlying assumptions can give rise to bias or have consequences for the interpretation of the estimated parameters. For this reason, it is important to check or at least to consider the assumptions underlying the methods used. Going from a model

$$g\{\mathrm{E}(Y_i | X_i, \mathbf{V}_i)\} = \mu + \beta X_i + \boldsymbol{\gamma}\mathbf{V}_i \tag{29}$$

for independent observations $i = 1, \ldots, M$ to a model

$$g\{\mathrm{E}(Y_{ij} | X_{ij}, \mathbf{V}_{ij}, \mathbf{W}_i)\} = \mu_i + \beta X_{ij} + \boldsymbol{\gamma}\mathbf{V}_{ij} \tag{30}$$

for clusters $i = 1, \ldots, M$ of individuals $j = 1, \ldots, n_i$, we also need to consider the potential associations between members of the same cluster. This has consequences for interpretation of results and inference from results. For instance, when using

conditional maximum likelihood (CML) estimation, is is important to ensure that there are no carry-over effects, i.e. that

$$Y_{ij} \perp\!\!\!\perp (X_{ij'}, \mathbf{V}_{ij'}) | X_{ij}, \mathbf{V}_{ij} \text{ for each } j' \neq j \ .$$

Violation of this assumption will lead to a biased CML estimate of β in (30) or to a different interpretation of β (Sjölander et al., 2016).

Another thing to consider with models for clustered data with cluster-specific intercepts is what is actually contained in $\mathbf{W}_i$. It is often argued that all shared confounding variables are adjusted for in CML estimation. The methods developed in Papers II and III will deal with the situation where $\mathbf{W}_i$ only consists of shared confounders for the association between exposures and outcomes. But it is often not explicit what $\mathbf{W}_i$ is, i.e. what is 'absorbed' into the cluster-specific intercepts μ_i. Do we also adjust for other shared factors such as shared mediators and shared colliders? The answer to this question does not have to do with the correctness of the assumed model, but with the parameter that is actually estimated in CML estimation. This is the subject of paper IV.

3 Aims

Using the Swedish Twin Registry or the Swedish Multi-generation Register it is possible to identify genetically related individuals in Sweden. Using the Swedish national identity numbers, unique to every individual in Sweden, it is possible collect a lot of data about those clusters. This gives medical researchers in Sweden opportunities to address research questions requiring adjustment for a large number of confounding factors, both measured confounding factors and shared non-measured confounding factors. This kind of data is uncommon outside of Scandinavia.

The large number of measured confounders also increases the risk of bias due to model misspecification. Even though there might exist estimators that can estimate the parameters in the main model without further distributional assumptions, such estimators tend to perform badly even in moderate sample sizes when there are a large numbers of potential confounding factors to adjust for (Robins and Ritov, 1997). For this reason, DR estimators are advantageous since they give the researcher two chances of getting a consistent estimate of the target parameter. Unfortunately, few DR estimators have been developed for parameters that quantifies within-cluster associations.

The aims of this thesis are threefold

- To implement existing DR estimators in an `R` package. This was the aim of Paper I.
- To develop DR estimators that can adjust for confounding by cluster and implement these method in an `R` package. This have been done in Papers II and III.
- To clarify some assumptions behind estimation methods for clustered data and to show their implications for interpretation of parameter estimates from twin and sibling studies. This was the subject of Paper IV.

4 Summary of papers

4.1 Paper I

Paper I is a self-contained theoretical introduction to DR estimation for independent data and to the `R` package `drgee` (Zetterqvist and Sjölander, 2015). The package is available at https://CRAN.R-project.org/package=drgee. Apart from DR estimation, the package also allows semi-parametric estimation of standard regression models, i.e. models of the type used in generalized linear models and in generalized estimating equations. The standard errors are estimated using a sandwich estimator (Stefanski and Boos, 2002). If a cluster-identifying variable is supplied, cluster-robust estimates of the standard errors are calculated instead.

4.1.1 Doubly robust G-estimation and G-estimation

In DR G-estimation (Robins, 2000b), we estimate an effect parameter β, quantifying an association between an exposure X and an outcome Y adjusted for a set of covariates $\mathbf{V}$ by combining an outcome model

$$g\{\mathrm{E}(Y|X,\mathbf{V})\} = \beta X + \boldsymbol{\gamma}\mathbf{V} \ , \tag{31}$$

where g is the identity or log link function, with a model for the distribution of X conditional on $\mathbf{V}$

$$h\{\mathrm{E}\{X|\mathbf{V}\}\} = \boldsymbol{\alpha}\mathbf{V} \ . \tag{32}$$

where h is some link function, not necessarily the same as g. The estimator can be formulated as a two-step estimator, where the nuisance parameter estimates $\hat{\boldsymbol{\gamma}}$ and $\hat{\boldsymbol{\alpha}}$ are obtained in the first step (e.g. using ML estimation) and β is estimated in the second step using $\hat{\boldsymbol{\gamma}}$ and $\hat{\boldsymbol{\alpha}}$. The obtained estimator of β is DR, i.e. it is consistent for the parameter β in the model

$$g\{\mathrm{E}(Y|X,\mathbf{V})\} - g\{\mathrm{E}(Y|X=0,\mathbf{V})\} = \beta X \ , \tag{33}$$

if either the 'outcome nuisance model'

$$g\{\mathrm{E}(Y|X=0,\mathbf{V})\} = \boldsymbol{\gamma}\mathbf{V} \tag{34}$$

or the 'exposure nuisance model' (32) is correctly specified, not necessarily both. The standard (non-DR) G-estimator (Robins et al., 1992) is obtained by setting the model (34) to 0.

4.1.2 Doubly robust estimation for logistic models

In DR estimation for logistic regression models where both outcome Y and exposure X are binary, we combine a prospective logistic regression model, e.g.

$$\text{logit}\{\Pr(Y = 1|X, \mathbf{V})\} = \beta X + \boldsymbol{\gamma}\mathbf{V} \ , \tag{35}$$

with a retrospective logistic regression model, e.g.

$$\text{logit}\{\Pr(X = 1|Y, \mathbf{V})\} = \beta Y + \boldsymbol{\alpha}\mathbf{V} \ . \tag{36}$$

The models (35) and (36) targets the same log odds ratio

$$\beta = \log\left\{\frac{\Pr(Y = 1, X = 1|\mathbf{V})\Pr(Y = 0, X = 0|\mathbf{V})}{\Pr(Y = 0, X = 1|\mathbf{V})\Pr(Y = 1, X = 0|\mathbf{V})}\right\} \tag{37}$$

but with different ways to adjust for $\mathbf{V}$. The prospective model (35) uses the outcome nuisance model

$$\text{logit}\{\Pr(Y = 1|X = 0, \mathbf{V})\} = \boldsymbol{\gamma}\mathbf{V} \tag{38}$$

and the retrospective model (35) uses the exposure nuisance model

$$\text{logit}\{\Pr(X = 1|Y = 0, \mathbf{V})\} = \boldsymbol{\alpha}\mathbf{V} \ . \tag{39}$$

We implement an estimator proposed by Tchetgen Tchetgen et al. (2010) which is DR, i.e. it is consistent for the log odds ratio β in (37) when at least one of the nuisance models (38) and (39) is correctly specified.

4.1.3 The `R` package `drgee`

The main function `drgee` lets the user supply two formulas, one for the outcome nuisance model and one for the exposure nuisance model. The exposure and outcome is inferred from left hand sides of the two formulas, respectively. There are also

arguments for the outcome link and the exposure link. Although the default is to use DR estimation, it is also possible to perform standard regression, G-estimation or retrospective logistic regression. In these cases only one of the nuisance models are used. The variance of the parameter estimates are calculated using a sandwich estimator (Stefanski and Boos, 2002). When a cluster-identifying variable is supplied, a cluster-robust variance is also estimated. It is also possible to model effect modification in the main model (33) by replacing $\boldsymbol{\beta}X$ with $\boldsymbol{\beta}\mathbf{Z}X$, where $\mathbf{Z} = \mathbf{Z}(\mathbf{V})$ is some function of $(\mathbf{V})$.

4.2 Paper II

In Paper II, we develop a DR estimator for within-cluster parameter, i.e. a DR estimator for a parameter β defined as

$$g\{E(Y_{ij}|X_{ij}, \mathbf{V}_{ij}, \mathbf{W}_i)\} - g\{E(Y_{ij}|X_{ij} = 0, \mathbf{V}_{ij}, \mathbf{W}_i)\} = \beta X_{ij} \tag{40}$$

where the outcome nuisance model have the form

$$g\{E(Y_{ij}|X_{ij} = 0, \mathbf{V}_{ij}, \mathbf{W}_i)\} = \mu_i + \boldsymbol{\gamma}\mathbf{V}_{ij} \ . \tag{41}$$

The method is based on CGEE, described in Section 2.5.4 and can also be seen as an extension of this methodology. For this reason, this estimator is only useful when g is the identity or log link. The estimator is constructed by augmenting the models (40) and (41) with a model

$$h\{E(X_{ij}|\mathbf{V}_{ij}, \mathbf{W}_i)\} = \nu_i + \boldsymbol{\alpha}\mathbf{V}_{ij}, \tag{42}$$

for the exposure. We show that the resulting estimator is DR, i.e. it is consistent for β in (40) when at least one of the nuisance models (41) and (42) is correctly specified, not necessarily both. To allow for effect modification of $\mathbf{V}_i$ and $\mathbf{W}_i$ we can replace βX_{ij} in (40) with $\boldsymbol{\beta}\mathbf{Z}_{ij}X_{ij}$, where $\mathbf{Z}_{ij} = \mathbf{Z}_{ij}(\mathbf{V}_i, \mathbf{W}_i)$ is some function of $\mathbf{V}_i$ and observed parts of $\mathbf{W}_i$. The nuisance models (41) and (42) can also be replaced by more general models that also includes second order or non-linear functions $\mathbf{V}_{ij}$.

In analogy to the DR G-estimator, an estimator based on an exposure model only

can be obtained by using the outcome nuisance model

$$g\{E(Y_{ij}|X_{ij}=0,\mathbf{V}_{ij},\mathbf{W}_i)\}=\mu_i\ ,$$

i.e. by setting $\boldsymbol{\gamma}$ to 0.

The nuisance models (41) and (42) can also be made more flexible by replacing $\boldsymbol{\gamma}\mathbf{V}_{ij}$ and $\boldsymbol{\alpha}\mathbf{V}_{ij}$ with more general functions of $\mathbf{V}_{ij}$ allowing non-linear functions of and interactions between elements of $\mathbf{V}_{ij}$.

The estimator is implemented in the `R` package `drgee`, with very similar syntax as for independent data.

Remark: The efficiency calculations found in Section S5 of the Supplementary Materials to Paper II were done by Stijn Vansteelandt.

4.3 Paper III

In study III, we develop a DR estimator for estimation using logistic regression models with cluster-specific intercepts where both exposure and outcomes are binary. As with logistic regression for independent data, there is also a 'prospective' and a 'retrospective' version of the conditional logistic regression. The 'prospective' conditional logistic regression model have the form

$$\text{logit}\{\Pr(Y_{ij}=1|X_{ij},\mathbf{V}_{ij},\mathbf{W}_i)\}=\beta X_{ij}+\mu_i+\boldsymbol{\gamma}\mathbf{V}_{ij} \tag{43}$$

and the 'retrospective' logistic regression model have the form

$$\text{logit}\{\Pr(X_{ij}=1|Y_{ij},\mathbf{V}_{ij},\mathbf{W}_i)\}=\beta Y_{ij}+\nu_i+\boldsymbol{\alpha}\mathbf{V}_{ij}\ . \tag{44}$$

Both versions share the log odds ratio

$$\text{logit}\{\Pr(Y_{ij}=1|X_{ij}=1,\mathbf{V}_{ij},\mathbf{W}_i)\}-\text{logit}\{\Pr(Y_{ij}=1|X_{ij}=0,\mathbf{V}_{ij},\mathbf{W}_i)\}=\beta\ , \tag{45}$$

but 'prospective' conditional logistic regression model uses the outcome nuisance model

$$\text{logit}\{\Pr(Y_{ij}=1|X_{ij}=0,\mathbf{V}_{ij},\mathbf{W}_i)\}=\mu_i+\boldsymbol{\gamma}\mathbf{V}_{ij} \tag{46}$$

and the 'retrospective' logistic regression model uses the exposure nuisance model

$$\text{logit}\{\Pr(X_{ij} = 1|Y_{ij} = 0, \mathbf{V}_{ij}, \mathbf{W}_i)\} = \nu_i + \boldsymbol{\alpha}\mathbf{V}_{ij} \ . \tag{47}$$

The DR estimator is consistent for β in (45) when at least one of the nuisance models (46) and (47) is correctly specified.

The estimator utilizes that fact that, for doubly discordant pairs, i.e. paired observations with $Y_{i1} + Y_{i2} = 1$ and $X_{i1} + X_{i2} = 1$, both models (43) and (44) can be rewritten as logistic models with common intercepts (Holford et al., 1978). The prospective logistic model (43) can be rewritten as

$$\text{logit}\{\Pr(Y_{i1} = 1|X_{i1}, Y_{i1}+Y_{i2} = 1, X_{i1}+X_{i2} = 1, \mathbf{V}_{ij}, \mathbf{W}_i)\} = 2\beta X_{i1}-\beta+\boldsymbol{\gamma}(\mathbf{V}_{i1}-\mathbf{V}_{i2})$$

and the retrospective logistic regression model (44)

$$\text{logit}\{\Pr(X_{i1} = 1|Y_{i1}, Y_{i1}+Y_{i2} = 1, X_{i1}+X_{i2} = 1, \mathbf{V}_{ij}, \mathbf{W}_i)\} = 2\beta Y_{i1}-\beta+\boldsymbol{\alpha}(\mathbf{V}_{i1}-\mathbf{V}_{i2}) \ .$$

Using these logistic forms we can utilize the DR estimator proposed by Tchetgen Tchetgen et al. (2010) and which was implemented in the `R` package `drgee` in Study I.

The method can be extended to also handle data with arbitrary cluster sizes, by considering all possible pairs within each cluster and by only using the doubly discordant pairs. The correlation between pairs from the same cluster is accounted for by using a cluster-robust sandwich estimator for the variance.

The estimator have been implemented in the `R` package `drgee`.

4.4 Paper IV

It is often argued that the 'within-cluster' associations are automatically adjusted for all factors that are shared within clusters. One motivation to use twin and sibling designs is that the estimated 'within-cluster' association parameter is adjusted for shared confounding. However, not all factors should be adjusted for if one wants to make causal inference. Adjusting for 'colliders', i.e. common effects of exposure and

outcomes, can introduce biased estimates of the causal parameters. Whether one should adjust for 'mediators', i.e. factors on the causal pathway between exposure and outcome, depends on the research question. What 'within-cluster' associations are actually adjusted for, is not explicit. As this has implications for the interpretation of results from twin and sibling studies, it is important to clarify this.

In twin and sibling studies, the assumed regression model is often vague regarding what is actually conditioned on. The vagueness stems partly from the fact that the shared factors are not observed and are not modeled individually. In this thesis, we have used $\mathbf{W}_i$ to refer to the shared factors that are adjusted for and assumed the regression model

$$g\{\mathrm{E}(Y_{ij}|X_{ij}, \mathbf{W}_i)\} = \mu_i + \beta X_{ij} \ . \tag{48}$$

In studies II and III, we assumed that $\mathbf{W}_i$ consists of the shared confounders $\mathbf{U}_i$. However, since there may also exist shared mediators $\mathbf{M}_i$ as well as shared colliders $\mathbf{C}_i$, it is natural to ask the question: *Which of the shared factors* $\mathbf{U}_i$, $\mathbf{M}_i$ *and* $\mathbf{C}_i$ *is the 'within-cluster' association adjusted for?* The answer to this question does not lie in the model itself, but in how β is estimated. When using CML to estimate β, there are two conditions that needs to be fulfilled in order to guarantee consistent estimation of β (Zetterqvist et al., 2016):

$$\text{A: } Y_{ij} \perp\!\!\!\perp Y_{ij'}|\mathbf{X}_i, \mathbf{W}_i$$
$$\text{B: } Y_{ij} \perp\!\!\!\perp X_{ij'}|X_{ij}, \mathbf{W}_i \text{ for } j \neq j'$$

By using DAGs, we show that assumptions A and B will be violated if $\mathbf{C}_i$ is *included in* $\mathbf{W}_i$ or if $\mathbf{U}_i$ or $\mathbf{M}_i$ is *omitted from* $\mathbf{W}_i$. Therefore, CML estimation will only consistently estimate β in the model

$$g\{\mathrm{E}(Y_{ij}|X_{ij}, \mathbf{U}_i, \mathbf{M}_i)\} = \mu_i + \beta X_{ij} \ , \tag{49}$$

even though there may also exist shared colliders $\mathbf{C}_i$.

It should be noted that the model (48) may be correctly specified even though $\mathbf{W}_i$ consists of another combination of the shared factors. However, the CML is only

guaranteed to be consistent for the β defined in (49). Even though assumptions A and B are not necessary conditions for consistent estimation of the target parameter, they are sufficient conditions (along with standard regularity conditions (see e.g. Andersen, 1970)). This means that if the CML estimator also is consistent for β^* in another model, e.g. for the model

$$g\{\mathrm{E}(Y_{ij}|X_{ij}, \mathbf{U}_i)\} = \mu_i + \beta X^*_{ij} \ , \tag{50}$$

we must have that $\beta^* = \beta$.

One important conclusion to draw from this result is that, the 'within-cluster' estimate that is obtained in twin and sibling studies can have a causal interpretation, since it is adjusted for shared confounders but not for shared mediators. Another conclusion is that the 'within-cluster' association is also adjusted for shared mediators $\mathbf{M}_i$. Therefore, β only measures the part of the effect that it is *not* mediated by $\mathbf{M}_i$. Thus, if most of the effect of X_{ij} on Y_{ij} is mediated by shared factors $\mathbf{M}_i$, absence of a significant 'within-cluster' association can not necessarily be used to support the claim of an absence of a causal effect.

5 Discussion

The Swedish total population registers together with the Swedish personal identity numbers is a great resource for epidemiological research, since the large amount of data enables adjustment for a large number of potential confounding variables (Ludvigsson et al., 2009, 2016).

Although the large number of available variables in the Swedish registers gives researchers opportunities to adjust for a large number of confounding factors it also increases the risk of bias due to model misspecification. Unfortunately, neither stratification nor nonparametric smoothing is feasible due to the curse of dimensionality (Robins and Ritov, 1997). For this reason, DR estimators are attractive in that they only require one of two models to be correctly specified (Robins et al., 2000; Robins and Rotnitzky, 2001), thus offering some protection against bias due model misspecification, while not being affected by the curse of dimensionality. Despite the advantages of DR estimators, they are still relatively seldom used in epidemiologic research. One obstacle has been the lack software implementing DR estimators. In study I, we have tried to remedy this by implementing DR G-estimation and DR logistic regression in the `R` package `drgee`. We have also tried to speed up the calculations when dealing with large datasets.

Two important resources for epidemiological research are the Swedish Multi-generation Register (Ekbom, 2011) and the Swedish Twin Registry (Pedersen et al., 2002; Lichtenstein et al., 2002, 2006), since they contain information about genetic relationships. Such information allows for adjustment of unobserved factors that are shared within clusters (e.g. twins, siblings, cousins). Apart from shared genetic factors, it is also possible to adjust for other factors, e.g. uterine environment (twins) and childhood environment (siblings), that are hard or impossible to measure.

In studies II and III, we have developed DR estimation methods that adjust for shared factors shared within clusters. More specifically, these estimation methods are robust against misspecification of the model for the influence of non-shared factors.

Despite the title of this thesis, the methods described in studies II, III and IV can also be used for general clustered data such as matched case-control data and matched cohort data. Another important application is when the aim is to measure a 'within-individual' association using longitudinal data. The estimate is then adjusted for factors that are constant within individuals over time.

Even though twin and sibling designs are advantageous for control of shared confounding, interpretation of results from such studies needs to be done with caution. One thing to keep in mind is that for non-linear association measures like the odds ratio and the hazard ratio are 'non-collapsible'. This means that the crude measure and the 'within-cluster' measure are *different measures* and may be different even in the absence of familial confounding. Both the crude odds ratio and the crude hazard ratio are closer to 1 compared to their 'within-cluster' counterparts. But even when the 'within-cluster' association measure is closer to 1, great caution is warranted when interpreting results from twin and sibling studies. For instance, Frisell et al. (2012) demonstrated that the 'within-cluster' association measures is more sensitive than the crude association measure to random errors in the measurements of the exposures. This will lead to measures of the 'within-cluster' association being closer to the null hypothesis than the crude measures, even in the absence of familial confounding. Another thing to keep in mind is that the methods used to estimate the 'within-cluster' parameters have other requirements on dependencies within the data. When there are carry-over effects, i.e. when the outcome of one member of the cluster is not independent of non-shared factors from another member of the cluster, the estimates may also be biased (Sjölander et al., 2016). In study IV, we note another consequence of these requirements: the 'within-cluster' measure is not only adjusted for shared confounders, it is also adjusted for shared mediators. This means that, whenever there are shared mediators, the crude association measure is a different measure that the 'within-cluster' association measure. Therefore, absence of a 'within cluster' association can not necessarily be interpreted as an indication of absence of an effect.

5.1 Limitations

The present work should be viewed in the light of some limitations.

First, even though DR estimators are consistent when at least one of two models is correct, the result gives no guidance about which model is correct. However, Robins and Rotnitzky (2001) noted that a comparison between the DR estimate and estimates based on one nuisance model only (e.g. GLM or G-estimation) can be useful as an informal goodness-of-fit test. If the DR estimate differs substantially from the two other estimates, one can conclude that both nuisance models are grossly misspecified and that all three estimates are biased.

Second, DR estimators are in general less efficient than ML estimators when the likelihood model is correctly specified. Thus the doubly robust property comes at a price. Further, the proposed estimators in studies II and III are not optimal in terms of efficiency, However, we have found no great loss in efficiency in our simulation studies.

Third, all models are to some extent misspecified. Thus, like all model-based estimators, DR estimators are generally biased in practice. In some scenarios (with all models misspecified), DR estimators have been found to perform worse than estimators based on one regression model only (Kang and Schafer, 2007; Waernbaum, 2012). However, in these situations, the target parameter was the average treatment effect and the DR estimator combined IPW-based estimation with standard regression followed by standardization. In this thesis, we focus on the treatment effect conditional on covariates. In general, the DR G-estimator is less sensitive to model misspecification than the IPW-based estimator and in some scenarios, it is even consistent under misspecification of both nuisance models (Vansteelandt et al., 2012).

Fourth, even if DR estimators are most often formulated in contexts of causal inference they are not guarantee valid causal inference, since even non-causal associations can be correctly specified and since there is no protection against misspecification of the main model.

5.2 Extensions

Except for special cases, the proposed estimator in studies II and III are not semiparametrically efficient, i.e. they are not optimal in terms of efficiency (Newey, 1990).

An interesting topic for future studies would be to find out if there exists alternative DR estimators of the same target parameters but with better efficiency.

In the studies I-III, we have used standard estimators for the nuisance parameters. As an alternative, it is possible to estimate $\boldsymbol{\gamma}$ and $\boldsymbol{\alpha}$ by directly minimizing the variance of the DR estimator, as described by Cao et al. (2009). Another option would be to estimate $\boldsymbol{\gamma}$ and $\boldsymbol{\alpha}$ by directly minimizing the bias of the DR estimator along the lines described by Vermeulen and Vansteelandt (2015). Doing this, one can avoid amplification of bias when both nuisance models are misspecified.

In project III, the nuisance model parameters are estimated using doubly discordant pairs only. Thus pairs that may contain information are discarded. As an alternative, the outcome nuisance parameter $\boldsymbol{\gamma}$ can be estimated by also including exposure concordant pairs that are discordant in both outcomes and nuisance factors $\mathbf{V}$. This would result in more 'precise' estimates of the nuisance parameter $\boldsymbol{\gamma}$. Similarly, using all exposure-discordant pairs can give more 'precise' estimates of the nuisance parameter $\boldsymbol{\alpha}$. However, it is not clear whether this would translate into a more efficient DR estimator of the target parameter β. It is well established that using estimated propensity scores instead of known propensity scores in IPW-based estimation results in more efficient estimation of the average treatment effect (Hirano et al., 2003; Brumback et al., 2010). The same is true for G-estimation when estimating the parameters for the exposure nuisance model (Robins et al., 1992). A topic for future studies would therefore to investigate, both theoretically and empirically, whether this is also true for the proposed DR estimator in study III.

A lot of time have been spend on the `R` package `drgee`. This is because we believe that offering software implementations of new methodology is crucial to make the methodology more attractive. There are still room for further refinement of the `drgee` package. For instance, an offset parameter would make it possible to estimate rate ratios using log-linear models in longitudinal studies. Another improvement would be to separate the DR estimation of the target parameter from the estimation of the nuisance parameters, to allow different methods for estimation of the nuisance

parameters or to use previously estimated nuisance parameters.

A potential future application of the proposed DR estimators would be in research on age-related and late-onset diseases where both life-style and genetic factors play important roles. As opposed to children and adolescents, adult and elderly are more likely to be discordant in life-style factors. The potentially large number of non-shared observed confounding factors in such studies would therefore make the risk of bias due to model misspecification large.

6 Acknowledgements

The research described in this thesis was carried out at the Department for Medical Epidemiology and Biostatistics (MEB) at Karolinska Institutet supported by grants from the Swedish Research Council. My time at MEB, both before and during my PhD studies, have been so rewarding on many levels that I cannot help feeling spoiled. I would like to thank the following people:

My main supervisor **Arvid Sjölander**. I feel honoured to have been your PhD student and I have learned so much from you. There are many things that I want to thank you for - for your time, for believing in me, for your patience with me, for your contagious enthusiasm, and many other things. I have always left our meetings with new energy after being reminded of the real reason why I am doing this - because it is fun! I could not have had a better supervisor.

My co-supervisor, boss, co-author and current head of the department **Paul Lichtenstein**. You were the first one from the department that I met and you have been the gate to the wonderful world of research. I am very grateful for the opportunity to participate in interesting research projects and the nice work environment that you help to create. I have learned a lot from working with you (and your guts).

My co-supervisor, co-author and former undergraduate supervisor **Henrik Larsson**. Thank you for believing in me and for directing me to Arvid. You introduced me to research and I learned a lot about research from you during my first time at MEB before I became a PhD student.

My co-supervisor, co-author and teacher **Yudi Pawitan**. I am grateful for your time when I got lost in the mysteries of statistics. I will always remember your hands-on perspective.

My mentor and co-author **Yasmina Molero Samuelson**. Glad to have you to discuss everything from science, statistics and career choices to animals and music.

My co-authors and collaborators **Stijn Vansteelandt** and **Karel Vermeulen**. Thank you for sharing your knowledge and for your kindness. It has been pleasure to work with you.

My R-, Emacs- and LaTeX-gurus **Alexander Ploner**, **Robert Karlsson**, **Henric**

Winell, **Mark Clements** and **Andreas Karlsson** for your invaluable support and for being such nice guys.

Marie Jansson, **Camilla Ahlqvist** and **Gunilla Nilsson Roos** for your kindness and for your patience with me. Without you around, there would be total chaos.

The former heads of the department **Henric Grönberg**, **Nancy Pedersen** and **Hans-Olov Adami**. Thank you for the stimulating research environment at MEB.

My ~~secretary~~ PhD colleague **Elisabeth Dahlqwist** for always making ~~fun of me~~ it fun to be at work.

My room-mates **Hannah Bower**, **Philippe Wagner** and **Maria Lantz** for putting up with me, all my stuff and my chatter.

Current and former colleagues for making MEB to such a nice workplace (in no particular order): **Bènèdicte Delcoigne, Gabriel Isheden, Anna Johansson, Thorgerdur Palsdottir, Flaminia Chiesa, Xingrong Liu, Peter Ström, Therese Andersson, Emilio Morales, Tong Gong, Qi Chen, Zheng Zhang, Alexander Viktorin, Jie Song, Anna Kähler, Sarah Bergen, Sara Ekberg, Erik Petterson, Dhany Saputra, Johanna Holm, Behrang Mahjani, Jonas F. Ludvigsson, Sandra Eloranta, Tomas Frisell, Ralf Kuja-Halkola, Sara Öberg, Agnieszka Butwicka, Linda Abrahamsson, Linda Halldner Henriksson, Hatef Darabi, Paul Dickman, Paul Lambert, Patrik Magnusson, Agnes Gothby Ohlsson, Rikard Öberg, Rino Bellocco, Alessandra Grotta, Keith Humphreys, Marie Reilly, Nelson Ndegwa Gichora, Amir Sariaslan, Juni Palmgren, Sven Sandin, Daniela Mariosa, Robert Szulkin, Myeongjee Lee, Henrik Olsson, Chen Suo, Maya Alsheh-Ali, Sara Fogelberg, Kathleen Bokenberger, Isabell Brikell, Carolyn Cesta, Malin Ericsson, Ida Karlsson, Andreas Jangmo, Shadi Azam, Dylan Williams, Annika Tillander, Caroline Weibull, Nghia Vu, Yi Lu** and many, many others...

Mom, Dad and my sister (with family) for your support and interest in what I am doing.

Cecilia, my love. Thank you for your help, for your support, for putting up with me and for just being around.

O, for making it worthwhile.

References

Andersen, E. B. (1970). Asymptotic properties of conditional maximum-likelihood estimators. *Journal of the Royal Statistical Society. Series B (Methodological)* **32,** 283–301.

Bang, H. and Robins, J. M. (2005). Doubly robust estimation in missing data and causal inference models. *Biometrics* **61,** 962–973.

Breslow, N. (1981). Odds ratio estimators when the data are sparse. *Biometrika* **68,** 73–84.

Breslow, N. and Powers, W. (1978). Are there two logistic regressions for retrospective studies? *Biometrics* **34,** 100–105.

Brumback, B., Dailey, A., Brumback, L., Livingston, M., and He, Z. (2010). Adjusting for confounding by cluster using generalized linear mixed models. *Statistics & Probability Letters* **96,** 1650–1654.

Cao, W., Tsiatis, A. A., and Davidian, M. (2009). Improving efficiency and robustness of the doubly robust estimator for a population mean with incomplete data. *Biometrika* **96,** 723–734.

Chen, H. Y. (2007). A semiparametric odds ratio model for measuring association. *Biometrics* **63,** 413–421.

Ekbom, A. (2011). The Swedish multi-generation register. *Methods in Biobanking* pages 215–220.

Frisell, T., Öberg, S., Kuja-Halkola, R., and Sjölander, A. (2012). Sibling comparison designs: Bias from non-shared confounders and measurement error. *Epidemiology* **23,** 713–720.

Goetgeluk, S. and Vansteelandt, S. (2008). Conditional generalized estimating equations for the analysis of clustered and longitudinal data. *Biometrics* **64,** 772–780.

Greenland, S. and Pearl, J. (2011). Adjustments and their consequencescollapsibility analysis using graphical models. *International Statistical Review* **79,** 401–426.

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

Hernán, M. A., Clayton, D., and Keiding, N. (2011). The Simpson's paradox unraveled. *International journal of epidemiology* **40,** 780–785.

Hirano, K., Imbens, G. W., and Ridder, G. (2003). Efficient estimation of average treatment effects using the estimated propensity score. *Econometrica* **71,** 1161–1189.

Holford, T., White, C., and Kelsey, J. (1978). Multivariate analysis for matched case-control studies. *American Journal of Epidemiology* **107,** 245–256.

Janes, H., Dominici, F., and Zeger, S. (2010). On quantifying the magnitude of confounding. *Biostatistics* **11,** 572–582.

Kang, J. D. and Schafer, J. L. (2007). Demystifying double robustness: A comparison of alternative strategies for estimating a population mean from incomplete data. *Statistical science* **22,** 523–539.

Liang, K.-Y. and Zeger, S. L. (1986). Longitudinal data analysis using generalized linear models. *Biometrika* **73,** 13–22.

Lichtenstein, P., Floderus, B., Svartengren, M., Svedberg, P., Pedersen, N. L., et al. (2002). The Swedish Twin Registry: a unique resource for clinical, epidemiological and genetic studies. *Journal of Internal Medicine* **252,** 184–205.

Lichtenstein, P., Halldner, L., Zetterqvist, J., Sjölander, A., Serlachius, E., Fazel, S., Långström, N., and Larsson, H. (2012). Medication for attention deficit–hyperactivity disorder and criminality. *New England Journal of Medicine* **367,** 2006–2014.

Lichtenstein, P., Sullivan, P. F., Cnattingius, S., Gatz, M., Johansson, S., Carlström, E., Björk, C., Svartengren, M., Wolk, A., Klareskog, L., et al. (2006). The Swedish Twin Registry in the third millennium: an update. *Twin Research and Human Genetics* **9,** 875–882.

Ludvigsson, J. F., Almqvist, C., Bonamy, A.-K. E., Ljung, R., Michaëlsson, K., Neovius, M., Stephansson, O., and Ye, W. (2016). Registers of the Swedish total population and their use in medical research. *European Journal of Epidemiology* **31,** 125–136.

Ludvigsson, J. F., Otterblad-Olausson, P., Pettersson, B. U., and Ekbom, A. (2009). The Swedish personal identity number: possibilities and pitfalls in healthcare and medical research. *European Journal of Epidemiology* **24,** 659–667.

McCullagh, P. and Nelder, J. A. (1989). *Generalized linear models*, volume 37. CRC press.

Neuhaus, J. M. and Jewell, N. P. (1993). A geometric approach to assess bias due to omitted covariates in generalized linear models. *Biometrika* **80,** 807–815.

Newey, W. K. (1990). Semiparametric efficiency bounds. *Journal of Applied Econometrics* **5,** 99–135.

Neyman, J. and Scott, E. (1948). Consistent estimates based on partially consistent observations. *Econometrica: Journal of the Econometric Society* **16,** 1–32.

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

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

Pedersen, N. L., Lichtenstein, P., and Svedberg, P. (2002). The Swedish Twin Registry in the third millennium. *Twin Research* **5,** 427–432.

Prentice, R. (1976). Use of the logistic model in retrospective studies. *Biometrics* pages 599–606.

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

Robins, J. M. (2000a). Marginal structural models versus structural nested models as tools for causal inference. In *Statistical models in epidemiology, the environment, and clinical trials*, pages 95–133. Springer.

Robins, J. M. (2000b). Robust estimation in sequentially ignorable missing data and causal inference models. In *Proceedings of the American Statistical Association*, volume 1999, pages 6–10.

Robins, J. M., Mark, S. D., and Newey, W. K. (1992). Estimating exposure effects by modelling the expectation of exposure conditional on confounders. *Biometrics* **48,** 479–495.

Robins, J. M. and Ritov, Y. (1997). Toward a curse of dimensionality appropriate(coda) asymptotic theory for semi-parametric models. *Statistics in Medicine* **16,** 285–319.

Robins, J. M. and Rotnitzky, A. (2001). Comment on 'Inference for semiparametric models: Some questions and an answer' by Bickel, P.J. and Kwon, J. *Statist. Sinica* **11,** 920–936.

Robins, J. M., Rotnitzky, A., and van der Laan, M. (2000). Comment on 'On profile likelihood' by Murphy, S.A. and van der Vaart, A.W. *Journal of the American Statistical Association* **95,** 477–482.

Scharfstein, D. O., Rotnitzky, A., and Robins, J. M. (1999). Adjusting for nonignorable drop-out using semiparametric nonresponse models. *Journal of the American Statistical Association* **94,** 1096–1120.

Sjölander, A., Frisell, T., Kuja-Halkola, R., Öberg, S., and Zetterqvist, J. (2016). Carryover effects in sibling comparison designs. *Epidemiology* **27,** 852–858.

Stefanski, L. A. and Boos, D. D. (2002). The calculus of M-estimation. *The American Statistician* **56,** 29–38.

Tchetgen Tchetgen, E. J., Robins, J. M., and Rotnitzky, A. (2010). On doubly robust estimation in a semiparametric odds ratio model. *Biometrika* **97,** 171–180.

van der Laan, M. J. and Robins, J. M. (2003). *Unified methods for censored longitudinal data and causality*. Springer Science & Business Media.

van der Vaart, A. W. (2000). *Asymptotic Statistics*. Cambridge University Press.

Vansteelandt, S., Bekaert, M., and Claeskens, G. (2012). On model selection and model misspecification in causal inference. *Statistical Methods in Medical Research* **21,** 7–30.

Vansteelandt, S., Joffe, M., et al. (2014). Structural nested models and G-estimation: The partially realized promise. *Statistical Science* **29,** 707–731.

Vermeulen, K. and Vansteelandt, S. (2015). Bias-reduced doubly robust estimation. *Journal of the American Statistical Association* **110,** 1024–1036.

Waernbaum, I. (2012). Model misspecification and robustness in causal inference: comparing matching with doubly robust estimation. *Statistics in Medicine* **31,** 1572–1581.

Zetterqvist, J. and Sjölander, A. (2015). Doubly robust estimation with the R package drgee. *Epidemiologic Methods* **4,** 69–86.

Zetterqvist, J., Vansteelandt, S., Pawitan, Y., and Sjölander, A. (2016). Doubly robust methods for handling confounding by cluster. *Biostatistics* **17,** 264–276.