Causal Inference and Case-Control Studies with Applications Related to Childhood Diabetes

Emma Persson

Department of Statistics
Umeå School of Business and Economics
Umeå University 2014
Contents

List of Papers i
Abstract ii
Preface iv

1 Introduction 1

2 The Potential Outcome Framework 2
  2.1 Identification of Causal Effects 3
  2.2 Covariate Selection 4
  2.3 Causal Parameters and Estimation 5

3 The Case-Control Design 6
  3.1 Independent Case-Control Design 7
  3.2 Matched Case-Control Design 8
  3.3 Parameters and Estimation 9

4 Type 1 Diabetes Mellitus (T1DM) 10
  4.1 The Swedish Childhood Diabetes Register 11

5 Summary of Papers 12
  5.1 Paper I 12
  5.2 Paper II 13
  5.3 Paper III 14
  5.4 Paper IV 15

6 Final Remarks and Further Research 16

Papers I-IV
List of Papers

The thesis is based on the following papers:


Paper II is reprinted with the kind permission of John Wiley & Sons.
Abstract

This thesis contributes to the research area of causal inference, where estimation of the effect of a treatment on an outcome of interest is the main objective. Some aspects of the estimation of average causal effects in observational studies in general, and case-control studies in particular, are explored.

An important part of estimating causal effects in an observational study is to control for covariates. The first paper of this thesis concerns the selection of minimal covariate sets sufficient for unconfoundedness of the treatment assignment. A data-driven implementation of two covariate selection algorithms is proposed and evaluated.

A common sampling scheme in epidemiology, and when investigating rare events, is the case-control design. In the second paper we study estimators of the marginal causal odds ratio in matched and independent case-control designs. Estimators that, under a logistic regression model, utilize information about the known prevalence of being a case is examined and compared through simulations.

The third paper investigates the particular situation where case-control sampled data is reused to estimate the effect of the case-defining event on an outcome of interest. The consequence of ignoring the design when estimating the average causal effect is discussed and a design-weighted matching estimator is proposed. The performance of the estimator is evaluated with simulation experiments, when matching on the covariates directly and when matching on the propensity score.

The last paper studies the effect of type 1 diabetes mellitus (T1DM) on school achievements using data from the Swedish Childhood Diabetes Register, a population-based incidence register. We apply theoretical results from the second and third papers in the estimation of the average causal effect within the T1DM population. A matching estimator that accounts for the matched case-control design is used.

Keywords: covariate selection; design-weighted estimation; marginal effect; matching; register study; treatment effect; type 1 diabetes mellitus
Preface

*Everything starts somewhere, although many physicists disagree.*

- Terry Pratchett, *Hogfather*

When I began my studies at Umeå University it was in physics, but after a while I found myself studying mathematical statistics. I am not entirely certain where I changed my mind, howbeit, here I am writing a PhD thesis in statistics. One thing that is certain is that I would not have started, continued or finished my PhD studies without the help and support of several people. I would like to take this opportunity to mention some of you and express my gratitude. It is largely because of you that I look back on these five years as a wonderful and fascinating experience, and not a gaping pit of despair.

First and foremost, I would like to thank my supervisor Ingeborg Waernbaum. I deeply appreciate your honesty, encouragement and guidance, without which this thesis would never have come to pass. Thank you for sharing this incredible experience with me and for being so generous with your time.

I would also like to thank my co-supervisor Xavier de Luna for all your valuable advise and feedback, and for asking difficult questions. To my second co-supervisor Gisela Dahlquist, your knowledge and enthusiasm is inspiring. Thank you both.

Next, I would like to thank everyone at the Department of Statistics who have made this the best of workplaces. My thanks to the Swedish Childhood Diabetes Study Group for fruitful and interesting discussions and for making the semiannual workshops a delight. Thank you Jenny and Maria, for good times both at work and outside of work. Maria, I partly blame you for this whole PhD thing.

I am also very grateful to all of my friends and family for your support and understanding. Mamma, thank you for feeding me! Daniel, thank you for always making sure that I am beer-wise well-irrigated. Thanks to Samantha and Maja for putting up with my incomprehensible babble. As Mr Data would have put it: *my mental pathways have*
become accustomed to your sensory input patterns.

Lastly, thank you Jens, for your constant support and above all for filling my life with laughter. Du är bäst!

Umeå, October 2014
Emma Persson
1 Introduction

In causal inference the main objective is to measure the effect of a cause, or treatment, on an outcome of interest. In this context, a treatment is defined in a broad sense and does not solely entail typical medical treatments. An educational achievement, a labor market program or a vaccine can also be seen as treatments under this definition (Holland 1986). For the evaluation of a medical treatment a randomized experiment is the most accepted and commonplace approach. In such experiments, where the effect of a treatment is studied in relation to no treatment (or another treatment), each unit under study is randomly assigned one of the two. Due to the randomization of the treatment assignment, we avoid systematic differences between treated and untreated, that may, in turn, affect the effect estimate.

However, there are situations where random experiments would be unfeasible, impractical or unethical to conduct. For research questions of this nature we need to rely on observational data. In a non-randomized (observational) study the treatment is not randomly assigned and there are likely characteristic differences between the treatment groups that are also related to the outcome. These differences make the effect of the treatment indistinguishable from the effect of these other variables. Therefore it is of importance to adjust (control) for these so called confounding variables in observational studies.

Following the framework of potential outcomes (Rubin 1974; Rubin 1977), which is sometimes referred to as the Rubin causal model (Holland 1986), this thesis focuses on certain issues concerning estimation of causal effects using observational data. There are many ways in which observational data may be collected. However, methods for statistical inference are generally based on the assumption of a random sample of units, or that observations are independent and identically distributed. It is not always possible or practical to collect a random sample, but choosing an appropriate sampling design is important since it determines the properties of the estimator subsequently used. When investigating a rare event, e.g., an uncommon disease, the case-control design is a suitable sampling design frequently applied.

This thesis consist of four papers contributing to the field of causal inference and case-control studies. Paper I focuses on selection of pretreatment variables to control for when estimating an average causal
effect in an observational study. The following three papers consider different aspects concerning estimation of average causal effects from case-control sampled data. In Paper II we study estimation of the marginal causal odds ratio in case-control designs. Assuming a logistic regression model, we compare different estimators that utilize information about the known prevalence of cases. The last two papers consider the special case when case-control sampled data is reused for the purpose of investigating research questions that classify the case-defining event as a treatment of interest. Paper III discusses the ramifications of ignoring the design and proposes a method of estimating the average causal effect in such situations. Lastly, in Paper IV we use matched case-control data from a population-based incidence register, to investigate the effect of type 1 diabetes mellitus (T1DM) on school achievements.

This thesis is structured as follows. It begins with an introduction to the causal framework in Section 2, followed by an overview of case-control studies in Section 3. In Section 4 the register data that allows for investigations related to T1DM is described. It is followed by a summary of the enclosed Papers I-IV and lastly some final remarks are given in Section 6.

2 The Potential Outcome Framework

The potential outcome framework was introduced by Neyman (1923) in the context of randomized experiments, and established by Rubin (1974, 1977) in observational studies as a framework for causal inference. In the following, the objective is to evaluate the effect of a binary treatment, $T$, on an outcome of interest. A unit under study is either exposed to the treatment, $T = 1$, or remain untreated, alternatively receives a control treatment, $T = 0$. The binary treatment implies that each unit has two potential outcomes. Let $Y_1$ denote the response variable if the unit is treated and $Y_0$ if the unit is untreated. Consider a sample of $n$ units drawn from a large and well-defined population. The causal effect for a specific unit $i$ (the individual causal effect) can be defined as the difference between the potential outcomes,

$$Y_{1i} - Y_{0i}, \quad i = 1, \ldots, n.$$ 

Since a unit can not both receive a treatment and remain untreated, only one of the potential outcomes is observed for each unit. Let $Y$
denote the observed outcome, where \( Y = TY_1 + (1 - T)Y_0 \). This issue is referred to as the \textit{fundamental problem of causal inference}, and implies that calculating the individual causal effect from the observed data is impossible (Holland 1986; Holland and Rubin 1988). However, instead of individual comparisons we may compare potential outcomes across units within a defined population. The main parameter of interest is often the population average causal effect (Imbens and Wooldridge 2009)

\[
\beta = E(Y_1 - Y_0). \tag{1}
\]

In some studies the population average causal effect of the treated,

\[
\gamma = E(Y_1 - Y_0 \mid T = 1), \tag{2}
\]

may be a more relevant effect measure, and in this thesis the principal focus is on \( \beta \) and \( \gamma \).

\section*{2.1 Identification of Causal Effects}

An average causal effect, such as (1) or (2), can not be identified without further assumptions. The first assumption, often not explicitly stated, called the \textit{stable unit treatment value assumption} (Rubin 1980), is needed for the definition of the potential outcomes and hence the causal effect. It states that there is only one version of the treatment and that a unit’s potential outcomes are not affected by the treatment assigned to other units.

Consider for a moment a randomized experiment, where we are studying a treatment that has been randomly assigned to individuals in a population. In such a situation (1) may be identified by observing that

\[
\beta = E(Y_1 \mid T = 1) - E(Y_0 \mid T = 0) = E(Y \mid T = 1) - E(Y \mid T = 0).
\]

Here, we may condition on \( T \) since the treatment assignment, by design, is independent of the potential outcomes, and the conditional expectations can, in turn, be estimated with the observed data.

In observational studies this independence will ordinarily not apply. However, average causal effects can still be identified using observational data, under some additional assumptions. Let \( X \) denote a set of pre-treatment variables, called covariates in the sequel, which are observed
for all units. Firstly, we assume unconfoundedness,

\[(Y_1, Y_0) \perp \perp T \mid X,\]

where \(\perp \perp\) signifies independence (Dawid 1979). This states that the potential outcomes are independent of the treatment given the covariates, and is sometimes referred to as no unmeasured confounding. Secondly, we have the assumption of overlapping distributions

\[0 < P(T = 1 \mid X) < 1.\]

This assumption states that all units have a treatment probability that is strictly between 0 and 1, i.e., there is a positive probability of receiving both treatments for all realizations of \(X\). In the causal inference literature, these two assumptions together are often referred to as strong ignorability (Rosenbaum and Rubin 1983). Subsequently, under strong ignorability, the average causal effects can be identified by noting that

\[
\beta = E[E(Y_1 - Y_0 \mid X)] = E[E(Y_1 \mid T = 1, X) - E(Y_0 \mid T = 0, X)]
\]

\[
= E[E(Y \mid T = 1, X) - E(Y \mid T = 0, X)].
\]

Thus, \(\beta\) can be estimated from observed data by first conditioning on \(X\).

### 2.2 Covariate Selection

In observational studies, the covariates play an important role in the estimation of causal effects. Above we see that identification of a causal parameter requires an observed set of covariates that satisfies both overlap and unconfoundedness (strong ignorability). The assumption of unconfoundedness is generally not testable using observed data, therefore we ordinarily rely on subject matter knowledge to assess its credibility. Naturally, if a set of covariates, \(X\), upholds unconfoundedness, it may not if a covariate is removed from \(X\). On the other hand, if a new covariate is added to the set, it may render other covariates superfluous. It is reasonable to assume that the condition holds if we have no unmeasured confounders, i.e., covariates that affect both the outcomes and the treatment, although there are some exceptions (Sjölander 2009). In practice this may require access to a large amount of background variables.

Having access to a lot of variables may tempt a researcher to include all of these in the conditioning set. However, including redundant
covariates have been shown to be less than optimal for some types of estimators, such as non-parametric estimators (de Luna, Waernbaum, and Richardson 2011). Therefore, choosing an appropriate $X$, such that unconfoundedness still holds, is of importance. The choice of covariates can be made in different ways, e.g., as part of model specification (Brookhart et al. 2006; Kelcey 2011; Vansteelandt et al. 2012) or by testing different possible sets under a priori assumptions of sufficient sets (Greenland, Pearl, and Robins 1999). An alternative is to use a model-free data-driven method of selecting covariates prior to the estimation of the causal effect, and such a method is explored further in Paper I.

2.3 Causal Parameters and Estimation

Depending on the research question, different causal parameters may be of interest. A conditional parameter can be helpful for a clinician when deciding on a medical treatment for a patient with particular characteristics. On the other hand, a marginal parameter can be used to assess the effect of a treatment in a population as a whole, something that is useful for policy makers. In contrast to the marginal causal parameters, (1) and (2), a conditional parameter may be the average causal effect in a subpopulation classified by certain values of $X$, e.g.,

$$\beta(x) = E(Y_1 - Y_0 \mid X = x).$$

Consequently, by conditioning on or averaging over $X$ a causal parameter is either conditional or marginal.

Irregardless of the targeted parameter, controlling for the covariates by conditioning on $X$ is essential in observational studies. An intuitive way of doing this is by stratifying the data with regards to $X$. However, if the dimension of the covariate vector increases or if $X$ includes continuous covariates, the number of strata needed to adequately control for the covariates increases as well, and the data may then prove insufficient. Consider an example where $X$ consists of ten binary covariates. All possible realizations of $X$ is thus $2^{10}$, yielding 1024 strata. Here, there may not be units exposed to opposite treatments in all strata.

Similar to stratification is the concept of matching (Stuart 2010). Here, units are matched to one or more units in the opposite treatment group but with the same or similar values on $X$. Thus, by averaging
the differences in outcomes of the matched pairs, or strata, of units we can obtain an estimate of the causal effect. Similar to the stratified approach, sparse data may be an issue for a matching estimator, where the inability to find a unit with the same value on $X$ may induce bias. However, a major advance was presented when Rosenbaum and Rubin (1983) introduced the propensity score, $P(T = 1 | X)$. Now, instead of matching on the covariates directly, the dimension of the covariate vector may be reduced by matching on the propensity score. There are a variety of matching estimators, see Stuart (2010) for a comprehensive overview, and in Paper I, III and IV we include estimators that matches one-to-one with replacement (Abadie and Imbens 2006).

Another class of estimators, under strong ignorability, relies on weighting the observed outcomes with the inverse of the propensity score (Hirano, Imbens, and Ridder 2003). Doubly robust estimators (Cao, Tsiatis, and Davidian 2009) are augmented inverse probability weighting estimators that incorporates a model for the outcome. For a review of estimation of causal effects under strong ignorability see Imbens (2004) and Imbens and Wooldridge (2009).

Other methods of controlling for confounding include regression methods. Often a parametric model for the conditional expectation of $Y$ given $T$ and $X$ is assumed, implying some assumptions on distributions or functional form for the variables involved. As an example of how regression can be applied in causal inference, consider a linear regression model for a continuous outcome that assumes a constant treatment effect. The coefficient in the regression model corresponding to the treatment is equal to (1) under unconfoundedness (Senn, Graf, and Caputo 2007). Linear and logistic regression are two common parametric approaches for continuous and binary outcomes, respectively.

3 The Case-Control Design

The case-control design is a possible sampling scheme for an observational study. It is a common design in epidemiology and often used when investigating the effect of an exposure (treatment) on a rare event or condition, e.g., a disease (Schlesselman 1982). In a case-control study, the response is the binary variable indicating if a unit has experienced the event in question (a case) or not (a control). After defining a case
population, the sampling of units is performed conditional on the response, i.e., cases and controls are sampled separately. Thus, in a case-control study we may compare a sample of individuals with the particular disease (cases) with a set of disease-free controls sampled from the population giving rise to the cases.

The design is an efficient alternative to a cohort study (Breslow 1996; Rothman, Greenland, and Lash 2008). In some situations, it can be the only feasible sampling design, since for rare diseases the size of a random sample needs to be very large in order to include sufficiently many cases for the subsequent analysis. Case-control studies are also valuable since they can be easier, less time consuming and more cost-efficient to perform than conventional prospective studies, such as a cohort study (Schlesselman 1982). The design is not immune to problems though and have received criticism in the past. Design issues include confounding, selection bias, measurement error and missing data. However, these problems are not solely interrelated to the case-control design, but are, in fact, problems that all observational studies are susceptible to (Breslow, 1996, 2005).

Cases and controls can be sampled in a number of ways. In this thesis, we consider designs where the cases are randomly sampled from a defined source population. A larger amount of cases will increase precision of the effect estimate, therefore it is not uncommon to select all cases if data on the whole population is available. Furthermore, incidence registers are not uncommon, and in Sweden there are a number of registers spanning the whole Swedish population that can be used for research purposes. The way in which controls are sampled may differ depending on cost and availability of data, see Rothman, Greenland, and Lash (2008) for possible alternatives. The case-control designs considered in this thesis select controls in two different ways: by random sampling and through matching on one or several variables. The designs are described in the following subsections.

### 3.1 Independent Case-Control Design

The independent case-control design is sometimes referred to as the unmatched case-control design. The design is perhaps the most standard case-control design and dates back to the 1950s (Breslow 2005). Here, the controls constitutes a random sample of units that have not expe-
rienced the event that defines the case. More formally, the units under study consist of \( n_1 \) units randomly sampled from the distribution of \((T, X)\) given \( Y = 1 \) and \( n_0 \) units randomly sampled from \((T, X)\) given \( Y = 0 \). This design is referred to as Design A in Paper II and III.

3.2 Matched Case-Control Design

In the matched case-control design, controls are matched to cases so that they have the same, or a similar value, on one or several variables, denoted by \( M \) in the following. More specifically, \( n_1 \) units are first randomly sampled from the distribution of \((T, X, M)\) given \( Y = 1 \), where \( M \) is, throughout this thesis, assumed to be a categorical variable. The source population is then stratified by \( M \), and for each selected case, a sample of \( J > 0 \) controls are randomly selected from the same stratum as the corresponding case. In this way, we obtain a sample with \( n_0 = Jn_1 \) controls, with the same sample distribution of \( M \) as the cases. This matched sampling scheme is called Design B in Paper II and III. Matching performed in the sampling stage of a study should not be confused with the matching that may be part of an analysis, e.g., the matching estimators described in Section 2.3.

The chosen matching variable is believed to be a confounder of the effect of the exposure on the outcome of interest, and the matching will ensure a balance of cases and controls across the levels of \( M \). The main advantage of the design, compared to the unmatched case-control study, is not to control for confounding, but that it can improve efficiency when estimating an exposure effect (Thompson, Kelsey, and Walter 1982; Smith and Day 1984; Stürmer and Brenner 2001).

The matched case-control design is a popular study design in epidemiology. Though it may increase efficiency it may also have less desirable consequences (Breslow 1982, 1996). Firstly, it is essential to account for the design in the analysis and this may prove complicated. In addition, if matching is not performed on a true confounder it may result in a loss of efficiency or an increase in bias. This situation is referred to as overmatching and include several possibilities. If a variable is related only to the treatment it need not be adjusted for and by matching on it we typically experience a loss in efficiency (Thomas and Greenland 1983; Kupper et al. 1981). The stronger the correlation between the matching variable and the treatment is the greater the loss of precision.
will be. A more serious scenario involves matching on a variable that is affected by the treatment and, in turn, affects the outcome (a mediator) or is affected by the outcome (a collider). In these situations, it will not be possible to obtain unbiased estimates of the exposure effect (Greenland and Neutra 1981; Vandenbroucke et al. 2007).

To ensure adequate power in the analysis in both independent and matched case-control studies we need to consider the amount of controls to be sampled. It is generally recommended to sample no more than four or five controls per case since minimal gains in power are usually achieved by increasing this ratio (Ury 1975). Though, in some situations, the matched case-control study, in particular, may benefit from an increased case to control ratio (Hennessy et al. 1999).

### 3.3 Parameters and Estimation

The case-control design is a biased sampling scheme due to the disproportionate number of cases/controls in the sample compared to the source population. It is imperative to take this into account in the analysis making appropriate adjustments for the design. In case-control studies, and in other studies where the response variable of interest is binary, we often see other measures of the effect, apart from (1), such as the risk ratio or odds ratio (Hernan 2004). These are measures of the same effect on different scales. The most predominant effect measure in case-control studies is the odds ratio and a reason for its popularity is that Cornfield (1951) showed that it is invariant under the case-control design. That is, the odds ratio expressed in terms of the response variable given the exposure is equal to the odds ratio expressed in terms of the exposure given the response. In addition, he showed that if the response prevalence (proportion of cases in the population) is small then the odds ratio approximates the risk ratio. Among the major methodological developments in case-control studies are the Mantel-Haenszel method (Mantel and Haenszel 1959) for adjusting for confounding by stratification, and perhaps the most commonly used method today, logistic regression (Prentice and Pyke 1979; Breslow and Day 1980; Scott and Wild 1986), see Breslow (1996) for an overview.

The marginal causal odds ratio is defined as

$$\theta_{OR} = \frac{P(Y_1 = 1)/P(Y_1 = 0)}{P(Y_0 = 1)/P(Y_0 = 0)}$$

(3)
and its conditional counterpart is

$$\theta_{\text{or}}(x) = \frac{P(Y_1 = 1 \mid X = x)/P(Y_1 = 0 \mid X = x)}{P(Y_0 = 1 \mid X = x)/P(Y_0 = 0 \mid X = x)}.$$ 

As with (1), the causal odds ratio may be conditional or marginal depending on if we condition on $X$ or not. Because of the non-linear nature of the parameter the conditional and marginal odds ratio may differ even without the presence of confounding. Greenland, Pearl, and Robins (1999) referred to this as non-collapsibility. For example, if a covariate is associated with the outcome but not with the treatment, and if there is an effect ($\theta_{\text{or}} \neq 1$), the conditional odds ratio will be farther from one than its marginal counterpart (Hauck, Neuhaus, Kalbfleisch, and Anderson 1991). Statistical developments have focused mainly on the estimation of the conditional odds ratio. However, some recent developments in the estimation of marginal causal odds ratios include estimators using an instrumental variable, i.e., a variable associated with the exposure but not the outcome (Vansteelandt, Bowden, Babanezhad, and Goetghebeur 2011). In addition, van der Laan (2008) has proposed case-control weighted targeted maximum likelihood estimators for (3), that update a conditional response model using a fitted model for the propensity score. In Paper II, we study estimation of the marginal causal odds ratio.

### 4 Type 1 Diabetes Mellitus (T1DM)

In the papers of this thesis we present applications related to T1DM. T1DM is an autoimmune disease characterized by the lack of insulin production, which is needed to regulate several metabolic pathways mirrored by, e.g., the glucose levels in the blood. The onset of T1DM typically occurs in childhood and the vast majority of children diagnosed with diabetes have type 1.

The incidence rate of T1DM is increasing worldwide and Sweden has the second highest reported. Among Swedish children, the incidence rate has approximately doubled over the last 30 years (Berhan et al. 2011). Predictions using the present trend anticipate that there will be a doubling of new cases of T1DM in European children 0-5 years old and that the prevalence of cases up to 15 years old will have risen by 70% in 2020 (Patterson et al. 2009).
4.1 The Swedish Childhood Diabetes Register

Using data from the Swedish Childhood Diabetes Register (SCDR), the applications in Paper I-IV address different research questions related to T1DM.

In Sweden all children that have been newly diagnosed with T1DM initially get treatment at a pediatric clinic. The clinics report to the SCDR their cases of T1DM, with informed consent from the parents. Thus, since July 1977, the SCDR has recorded incident cases of T1DM, younger than 15 years of age. As of 2010, cases are identified through the Swedish National Drug Register, which contains information on pharmaceutical drugs dispensed through Swedish pharmacies. Children that have, for the first time, collected their insulin prescription (at least twice) are reported to the SCDR. The register has a high level of coverage (96-99%) (Nyström et al. 1990; Dahlquist et al. 1985) and presently includes approximately 17,000 individuals with childhood onset of T1DM.

The SCDR also includes a control group of non-diabetic individuals from the Swedish population. For each T1DM individual, four controls are sampled by Statistics Sweden and matched by year of birth and municipality of residence at the time of the T1DM diagnosis. Thus, the SCDR have a matched case-control design.

Using the unique Swedish personal identification number (PIN, or personnummer in Swedish), the SCDR has also been linked at the individual level to other population registers, giving access to a large amount of background variables. The linking procedures are performed at Statistics Sweden and the National Board of Health and Welfare and data are sent back coded without a code key as permitted by the Regional Research Ethics Board in Umeå. By linkage to the Multi-Generation Register (MGR) information on siblings and parents is also retrieved. The MGR includes individuals born after 1932 that have at some point been registered in Sweden after 1961, and their biological and adoptive parents. In the applications of this thesis we adjust for parental variables and this is made possible through the MGR which allows biological parents of the cases and controls to be identified.

Another population register is the Longitudinal Integration Database for Health Insurance and Labour Market Studies (LISA) which is maintained by Statistics Sweden and includes the Swedish population 16
years of age and older. LISA is updated annually since 1990 and contains information from the labor market and the educational and social sector. In this thesis, we use variables on income, unemployment benefits, civil status and level of education, and all of these stem from LISA.

In addition, in Paper I and IV, we utilize information from the Swedish Register of Education to gain information on school achievements. This register has information on schooling, such as grades, from 1985 and is updated yearly. Lastly, medical information on cases and controls is gathered from the Inpatient Register, the Swedish Medical Birth Register and the Swedish National Drug Register maintained by the National Board of Health and Welfare.

5 Summary of Papers

5.1 Paper I

This paper, entitled *Data-driven Algorithms for Dimension Reduction in Causal Inference*, concerns the selection of optimal sets of covariates for the estimation of average causal effects. In an observational study the causal effect of the treatment may be confounded, however, it can be identified under the assumption of unconfoundedness. In practice, a researcher will need to select a set of covariates that is believed to uphold this assumption using subject matter knowledge. Large databases and population registers containing many covariates may be available to the researcher, giving access to several set options. There may be multiple covariates on income, employment and education, all to some degree measuring socioeconomic status. The researcher may therefore be tempted to include all of these to ensure that unconfoundedness holds. However, including redundant covariates (not true confounders) in the analysis targeting an average causal effect may affect bias (Abadie and Imbens 2006, theorem 1) and efficiency (Hahn 2004; de Luna, Waernbaum, and Richardson 2011, proposition 5 & 6) of nonparametric estimators, negatively.

Under the assumption that an initial set of covariates, $X$, upholds unconfoundedness, de Luna, Waernbaum, and Richardson (2011) introduced algorithms for selection of minimal subsets of covariates such that the treatment and the potential outcomes are independent given these sets. The algorithms define subsets of $X$ by conditional independence
statements and in this paper we propose a data-driven implementation of the algorithms by using model-free tests. Two approaches, subject to different conditions, are suggested. When $X$ contains continuous covariates we use marginal co-ordinate hypothesis (MCH) tests (Cook 2004; Li, Cook, and Nachtsheim 2005) from the theory of sufficient dimension reduction in regression (Cook 1994, 1996). In the case where $X$ includes discrete covariates, or a mix of discrete and continuous covariates, a method based on kernel smoothing (Hall, Racine, and Li 2004; Hall, Li, and Racine 2007a, 2007b; Li, Racine, and Wooldridge 2009) is used to find the subsets.

The implemented algorithms’ abilities to select the targeted subsets are evaluated in simulations, and the impact of the dimension reduction on the performance of a matching estimator of the average causal effect (Abadie and Imbens 2006) is also highlighted. The simulation study shows that the algorithms, for both methods of testing conditional independence, have a high success rate in selecting subsets that uphold unconfoundedness.

Finally, an application using data from the SCDR is presented. The algorithms are used to select subsets for confounder control in the analysis investigating the effect of low compulsory school grades on acute complications of T1DM. Here, using the mean grade as a proxy for the individual’s ability to self-regulate the disease, we find a significant effect that individuals with low mean grades are more likely to be hospitalized with acute complications of T1DM.

5.2 Paper II

In case-control studies where the objective is to estimate the effect of a treatment on the outcome of interest, two issues need to be addressed: confounding due to the fact that the exposure under study is not randomized, and accounting for the sampling design to avoid bias. This implies that, for an empirical researcher, choosing an appropriate method and controlling for the relevant covariates are imperative. Theoretical development have been focused on estimation of the conditional odds ratio, however, marginal effect measures, which express the causal effect at the population level, may be of greater relevance for public health.

In Paper II, entitled *Estimating a marginal causal odds ratio in a case-control design: analyzing the effect of low birth weight on the risk of*
type 1 diabetes mellitus, we study some estimators of the marginal causal odds ratio in matched and unmatched case-control designs. Throughout the paper we assume that the prevalence of being a case, $P(Y = 1)$, is known. The estimators under study have in common that they all adjust for both confounding and the case-control design. A logistic regression model for the observed outcome conditional on the treatment and covariates is assumed, and the prevalence is used to adjust the model either with intercept adjustment (Anderson 1972; Prentice and Pyke 1979) or weighted maximum likelihood (Scott and Wild 1986; van der Laan 2008). Based on these approaches, two estimators of the marginal causal odds ratio are constructed and approximations of their variances are derived using the delta method. In addition, we compare them to case-control weighted targeted maximum likelihood estimators (TMLE) (van der Laan 2008) through simulations.

As an illustrating example we apply the estimators described in the paper to data from the SCDR, analyzing the effect of low birth weight on the risk of T1DM. No significant marginal effect of low birth weight on T1DM was found. However, we show that neglecting the confounding or considering the conditional estimate of the odds ratio result in a different conclusion than when considering the marginal estimate that adjusts for confounding.

5.3 Paper III

The popularity of the case-control study, when investigating the effect of a risk factor on a certain disease, implies that there may be an abundance of case-control sampled data available in the scientific community. In addition, it is not unusual for other research questions concerning the disease to arise, e.g., what is the effect of the disease on some outcome later in life? Now, can we use the resource of the already sampled case-control data to investigate such a question, and in that case, how?

Paper III, *Estimating Marginal Causal Effects in a Secondary Analysis of Case-Control Data*, investigates the particular situation where matched or unmatched case-control data is used in a secondary analysis to evaluate the effect of the case-defining event on later outcomes. Here, the role of the event has changed from a response variable (in the initial study) to a treatment of interest. The paper focuses on adjustment for the sampling design when estimating a population average causal
effect. For a general class of estimators, we show the components of the bias that results from ignoring the sampling scheme and demonstrate a design-weighted matching estimator based on the estimator in Abadie and Imbens (2006). The proposed design-weighted matching estimator utilizes information about the prevalence of being a case.

Since the introduction of the propensity score by Rosenbaum and Rubin (1983), it has become common for matching estimators to use the estimated propensity score. In Paper III, we also discuss the consequences of the sampling design for the estimation of the propensity score. In simulations, the finite sample properties of the proposed estimator is investigated when matching on the covariates directly, and the true and estimated propensity scores. The results suggest that the estimator gives reliable inference for moderate sample sizes.

Lastly, using data from the SCDR, we study the effect of T1DM on the use of anti-depressant medication, and the results indicate no significant marginal effect. However, the average causal effect of the treated is 0.02, indicating an increased risk of using antidepressant medication for the T1DM population.

5.4 Paper IV

This paper, entitled The effects of type 1 diabetes mellitus on school achievements: A Swedish study using register data, investigates the impact of T1DM on educational achievements among children born during the 1980s and early 1990s. To our knowledge, this is the first study investigating the impact of T1DM on school achievements among individuals that are born during the 1980s and afterwards. These individuals have, throughout their childhood, had access to many of the essential treatment improvements seen over the last decades.

Contrary to the negative health effects of T1DM, little is known about how the disease impacts other aspects of a child’s life. Several characteristics of T1DM may affect the individual’s school achievements and level of education. Data from the SCDR is used to analyze the conditional and marginal effect of T1DM on final school grades from compulsory and upper secondary school. The conditional effect is estimated assuming a linear regression model, and the average causal effect within the T1DM group is estimated using a matching estimator.

The main results from the study show slightly lower grades in both
compulsory and upper secondary school among children diagnosed with T1DM. The small negative effect remains when controlling for demographic and socioeconomic factors such as sex, age and maternal educational level.

6 Final Remarks and Further Research

This thesis contributes to the literature on causal inference and case-control studies. Some aspects concerning the estimation of average causal effects are explored. When modeling a causal effect, we may condition on covariates in order to study the conditional causal effect of a treatment on an outcome given some set of characteristics (covariate values). This corresponds to the traditional methods used in case-control studies. However, within the causal framework an average causal effect may be more relevant. The estimation of an average causal effect involves conditioning on covariates in order to adjust for confounding, and subsequently marginalizing over the covariate distribution.

In Paper I we present data-driven algorithms for the selection of covariate subsets sufficient for estimating average causal effects. In the case where discrete covariates are present, a method based on kernel smoothing is used to reduce the dimension of the covariate vector, and threshold values for the bandwidths have to be specified by the user. In the simulations, binary covariates showed more sensitivity to the choice of bandwidth than continuous covariates, and further exploration into how to choose the thresholds is needed, e.g., through sensitivity analysis.

In Paper II, we construct two estimators of the marginal causal odds ratio under case-control sampled data, and derive approximations of their variances using the delta method. An alternative method appropriate for variance estimation is by means of the influence function. Also, we study the TMLE estimator (van der Laan 2008) which utilizes the estimated propensity scores to update the outcome model. Though some truncation of the propensity scores is recommended for computational purposes, further investigations into suitable truncation options and their consequences in terms of identification are of interest.

In this thesis we also show that average causal effects can be estimated from case-control data when the case-defining event is considered the exposure of interest. We show, in simulations, that the
design-weighted matching estimator in Paper III, performs well in finite samples. Preliminary results show that a central limit theorem for the design-weighted matching estimator can be applied. Although, further work is needed for verification of certain conditions.

In Paper III, we also discuss estimation of the propensity score in a secondary analysis of case-control data. This situation, and any situation where the sampling is dependent on the treatment variable, implies that the data is ”outcome-dependent” when modeling the propensity score. Thus, if the sampling design, e.g., the case-control design, is not taken into account the estimated propensity score will be biased. Although both stratified (on treatment) samples and propensity score methods are common in practice, this is not something that is widely recognized, see Dehejia and Wahba (2002) for an example. However, we show that there are situations where the biased propensity score results in a balancing score (Rosenbaum and Rubin 1983) which, in turn, can be used for confounder control in the analysis.

References


