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

Method developments for the attributable fraction in causal inference

Elisabeth Dahlqwist

Stockholm 2019

All previously published papers and images were reproduced with permission from the publishers.

Published by Karolinska Institutet.
Printed by E-Print AB 2019.

© Elisabeth Dahlqwist, 2019

ISBN 978-91-7831-422-5

Institutionen för Medicinsk Epidemiologi och Biostatistik

Method developments for the attributable fraction in causal inference

AKADEMISK AVHANDLING

som för avläggande av medicine doktorsexamen vid Karolinska Institutet offentligen försvaras i hörsal Petérn, Nobels väg 12 B, Karolinska Institutet, Solna

Fredagen den 24 maj 2019, kl 09.00

av

Elisabeth Dahlqwist

Huvudhandledare:
Docent Arvid Sjölander
Karolinska Institutet
Inst. för Medicinsk Epidemiologi och Biostatistik

Bihandledare:
Professor Yudi Pawitan
Karolinska Institutet
Inst. för Medicinsk Epidemiologi och Biostatistik

Fakultetsopponent:
Professor Xavier de Luna
Umeå Universitet
Handelshögskolan, enheten för Statistik

Betygsnämnd:
Docent Liisa Byberg
Uppsala Universitet
Inst. för kirurgiska vetenskaper, Epihubben

Professor Paul Lambert
University of Leicester
Department of Health Sciences
samt
Karolinska Institutet
Inst. för Medicinsk Epidemiologi och Biostatistik

Docent Frank Miller
Stockholms Universitet
Statistiska Institutionen

Stockholm 2019

To all the people who inspired and supported me.

Abstract

In public health and policy making, understanding the overall impact of an intervention is of essential importance. A way to quantify the disease burden due to some risk factor is by the attributable fraction (AF). The AF is a measure of the proportion of some disease that could be prevented if all would have been unexposed to the risk factor of interest. From the definition of the AF, it is a causal parameter and in order to achieve a causal interpretation of the AF estimate, we have to tackle the challenges of estimating causal effects in observational data. One of them is the problem of confounding, which may cause the researcher to confuse a spurious correlation with a causal effect.

In this work, we stress the importance of using model-based adjustment to estimate the AF and develop novel methods for AF estimation. In **project I** we implemented methods for AF estimation for cross-sectional, case-control (matched and unmatched) and cohort study designs in the statistical software R by the package AF. The package serves as a platform for the novel methods of AF estimation developed in **project II-IV**.

While **project I** focuses on estimation methods for the AF that rely on the fact that all confounders, sufficient for confounding control, are measured, researchers often face the problem with unmeasured confounding. In some situations, we may have access to clusters that share these unmeasured confounders. Thus, clustered data can be used to adjust for cluster-shared unmeasured confounding. In **project II** we develop a method that enables estimation of the AF, as a function of time, and adjusts for cluster-shared unmeasured confounders.

In practice, confounders may be unmeasured, but not shared within clusters, or we may lack access to clustered data. One remedy is to use an instrumental variable to mimic a randomized controlled trial and estimate the causal effect. In **project IV**, we developed a method for AF estimation based on instrumental variable analysis.

Genetics play an important role in the disease development and the concept of heritability, i.e. the variation in a trait explained by genetic factors, is often used to quantify the role of genetics. However, heritability does not convey any information on the population impact of some disease due to genetics. In **project III** we show how the AF can be conceptualized for complex traits, with the overall genetic risk as the exposure, and how heritability and the AF are formally related.

List of scientific papers

I. Elisabeth Dahlqwist, Johan Zetterqvist, Yudi Pawitan and Arvid Sjölander. Model-based estimation of the attributable fraction for cross-sectional, case-control and cohort studies using the R package AF. *European Journal of Epidemiology* 2016; **31**: 575-582.

II. Elisabeth Dahlqwist, Yudi Pawitan and Arvid Sjölander. Regression standardization and attributable fraction estimation with between-within frailty models for clustered survival data. *Statistical methods in medical research* 2017; **28(2)**: 462-485.

III. Elisabeth Dahlqwist, Patrik KE Magnusson, Yudi Pawitan and Arvid Sjölander. On the relationship between the heritability and the attributable fraction. *Human Genetics* 2019 (epub ahead of print).

IV. Elisabeth Dahlqwist, Zoltán Kutalik and Arvid Sjölander. Using Instrumental Variables to estimate the attributable fraction. (*Submitted*)

These articles are referred to by their roman numerals throughout, and are presented in full at the end of this thesis.

The studies in this thesis were supported by Vetenskapsrådet.

Contents

List of abbreviations and mathematical notations

ADHD	Attention deficit hyperactivity disorder
AF	Attributable fraction
BW	Between-within
CHD	Coronary heart disease
DAG	Directed acyclic graph
$\mathrm{E}(\cdot)$	Expected value
$\exp(\cdot)$	Exponential function
G	Denotes instrumental variable or genetic exposure
GPS	Genome-wide polygenic score
GWAS	Genome-wide association study
IV	Instrumental variable
L	Liability
$\lambda(\cdot)$	Hazard function
LATE	Local average treatment effect
$\log(\cdot)$	Base-e log (or the natural logarithm)
MR	Mendelian randomization
NEM	No effect modification
OR	Odds ratio
PH	Proportional Hazard
$\Pr(\cdot)$	Probability
RCT	Randomized controlled trial
RR	Risk ratio
$S(\cdot)$	Survival function
SMM	Structural mean model
TS	Two stage
TSLS	Two stage Least Square
UKBB	United Kingdom biobank
U	Denotes unmeasured confounding or frailty
X	Generally denotes the exposure
Y	Generally denotes the outcome
Z	Generally denotes measured confounder/confounders

1 Introduction

We live in a world surrounded by a large variation of factors that may affect our lives and health. The understanding of how environmental or genetic factors influence us is necessary when designing public health interventions. For this purpose, the task of distinguishing a spurious correlation from a causal relationship is essential, but far from trivial in practice.

For example, based on observational data in New York, US, smoking was expected to cause lung cancer [1]. These discoveries were, however, not sufficient as evidence for declaring that smoking is a causal risk factor for lung cancer. The tobacco industry argued that the observed association may be due to genetic factors that increase the risk for becoming a smoker as well as getting lung cancer, i.e. the observed association was driven by a common cause, a confounder. By the use of different study designs and experimental evidence in animal studies, smoking could be proven to be a risk factor for lung cancer [2]. The process that lead to smoking finally being accepted as a risk factor for lung cancer illustrates the challenge in proving causal effects in epidemiological practice [3].

The gold standard for reducing the problem of confounding, and for proving causal effects, is to randomize the exposure of interest in a group, representative of the population of interest, i.e. a randomized controlled trial (RCT). However, in many situations RCTs are not feasible for practical and ethical reasons, for example, when we expect the exposure to be harmful, such as the example with smoking. Observational data is often used in those situations, or to complement the results from an RCT. Due to the problem of confounding, researchers aiming to estimate causal effects in observational data are faced with some great challenges [4], such as the problem with unmeasured confounding.

In epidemiology, the estimation of causal effects has traditionally been addressed by the Bradford-Hill criterias [5], which are a set of practical guidelines for the evidence needed in order to establish a causal effect. As a complement, the statistical field of causal inference has developed a formal framework for conceptualizing and estimating causal effects. This framework incorporates a definition of the causal parameter, statistical methods for consistent estimation and the conditions for parameter identification. What we today refer to as causal inference was initially formulated by Donald Rubin in the 1970s and later developed by James Robins, Judea Pearl, and Miguel Hernán, among others [6].

A special feature of epidemiology is the use of different measures to describe the risk of some outcome, e.g. a disease or medical state due to some exposure [5]. The risk ratio (RR) estimates the relative *risk* of some outcome between two groups, e.g. exposed versus unexposed, while the odds ratio (OR) estimates the relative *odds* of the outcome between two groups.

Even though measures such as the RR and OR help us to answer questions regarding the effect size of some risk factor, they are relative measures and do not give information about the absolute risk and population impact. In public health, an understanding of the population impact of some risk factor is especially important since it may aid in designing efficient interventions. One measure that estimates the proportion of some outcome that is attributable to some risk factor is the attributable fraction (AF).

1.1 The attributable fraction

The AF is a population-specific measure of the proportion of preventable outcomes, e.g. disease cases, had all subjects in the population been unexposed to an exposure of interest. One aspect that has made the AF popular in epidemiology and public health is that it quantifies the exposure-outcome relationship by taking the exposure prevalence into account. The AF was first used in the 1950's by Morton Levin in a study of the relationship between smoking and lung cancer [1] and later on, Macmahon and Pugh (1970) [7] defined the AF in Eq. (1.1) for a binary outcome Y and a binary exposure X

$$\text{AF}_{naive} = 1 - \frac{\Pr(Y = 1 \mid X = 0)}{\Pr(Y = 1)}, \tag{1.1}$$

where $\Pr(Y = 1 \mid X = 0)$ is the outcome prevalence among unexposed and $\Pr(Y = 1)$ is the overall outcome prevalence in the population. Reading Levine (1953) [1], it is clear that the AF_{naive} was intended as a causal measure, where the outcome prevalence among unexposed $\Pr(Y = 1 \mid X = 0)$ serves as a proxy for the outcome prevalence had everyone been unexposed (with exposure status $X = 0$). In causal inference, this quantity is called the counterfactual outcome prevalence, denoted as $\Pr(Y_0 = 1)$.

Whether $\Pr(Y = 1 \mid X = 0)$ equals $\Pr(Y_0 = 1)$ will depend on the level of confounding and in general, they will not be exchangeable in observational data. For example, if lifestyle factors or socioeconomic background are confounders of the association between smoking and lung cancer, the lung cancer prevalence among non-smokers may not equal the lung cancer prevalence, if everyone in the population had been non-smokers.

In causal inference, directed acyclic graphs (DAGs) are often used to graphically present the problem of confounding [8]. Let $\mathbf{Z}$ denote a set of confounders. Figure 1.1 is a DAG that describes the confounding by $\mathbf{Z}$ of the causal relationship between exposure status X and outcome Y.

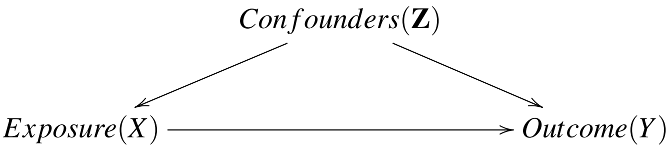

Figure 1.1: DAG of the causal relationship between exposure status X, outcome Y and confounders Z.

Thus, in the presence of confounding, the definition of the AF in Eq. (1.1) cannot have a causal interpretation. A definition of the AF which corresponds to a causal interpretation of the AF, also in observational data, is given in Sjölander and Vansteelandt (2011) [9],

$$\mathrm{AF} = 1 - \frac{\Pr(Y_0 = 1)}{\Pr(Y = 1)}. \tag{1.2}$$

However, if $\mathbf{Z}$ contains all confounders sufficient for confounding control, the equality in Eq. (1.3) holds.

$$\Pr(Y_0 = 1) = \mathrm{E}_{\mathbf{Z}}\{\Pr(Y = 1 \mid X = 0, \mathbf{Z})\}, \tag{1.3}$$

and the counterfactual outcome prevalence $\Pr(Y_0 = 1)$ can be estimated by adjusting for the confounders $\mathbf{Z}$.

Confounding adjustment can be made in different ways for the AF [10, 11] and different estimation strategies of $\mathrm{E}_{\mathbf{Z}}\{\Pr(Y = 1 \mid X = 0, \mathbf{Z})\}$ in Eq. (1.3), has been proposed for cross-sectional and case-control study designs [9, 12, 13].

The AF has also been defined for time-to-event outcomes [14–16]. Let T be the time-to-event of interest, the AF function is defined as

$$\mathrm{AF}(t) = 1 - \frac{\Pr(T_0 \leq t)}{\Pr(\mathrm{T} \leq t)}, \tag{1.4}$$

where $\Pr(T \leq t)$ is the factual probability of an event at or before time t, and $\Pr(T_0 \leq t)$ is the counterfactual probability of an event at or before time t, had the exposure been eliminated for everyone at baseline.

2 Problems and aims

There are several generalizations of the AF for multiple exposures [13, 18, 19], sequential mediation [20, 21], multiple exposures levels [18, 22–24] and multi-state models with time-varying exposures [25, 26], which allow for a more generalized modelling of the AF. In contrast to these methods, the focus of this thesis is on how to obtain a causal interpretation of the AF estimate by using the study design and facilitate their use by software implementation.

Depending on the study design, we may use different estimation strategies to estimate the AF [11]. For the standard study designs in epidemiology for observational data, i.e. cross-sectional, case-control and cohort, the theory for model-based estimations has been developed. At the beginning of 2015, no statistical software for model-based estimation of the AF in different sampling designs was available. Thus, the aim with **project I** was to create a package in the statistical software R [27] that covers model-based estimation of the AF for cross-sectional, case-control and cohort sampling designs. This package also served as a platform for the software implementations of **projects II-IV**.

A problem when estimating causal effects is the assumption of no unmeasured confounding. This assumption is rarely plausible in observational data but if the unmeasured confounders are shared within clusters, we can adjust for the cluster-shared confounders by statistical methods that condition on the cluster. For time-to-event outcomes, the stratified Cox proportional hazard (PH) model is used for this purpose. A limitation with the stratified Cox PH model is that we cannot estimate absolute probabilities. In **project II** we develop an estimation strategy for the AF for clustered time-to-event data by using the 'between-within' (BW) frailty model.

In general, we may not have access to clustered data, or clusters may not contain all unmeasured confounders of interest. Another remedy to avoid the problem with unmeasured confounding is to use an instrumental variable (IV) to mimic an RCT. In **project IV**, we develop a potentially confounding robust estimation strategy of the AF based on IV analysis.

The AF is mainly used for environmental exposures but genetic risk factors also play an important role in the disease development. The traditional ways to use single, or a set of risk increasing genetic variants to estimate the AF, cannot capture the overall genetic disease risk for complex diseases. In **project III** we investigate how the AF, with the overall genetic risk as the exposure, can be conceptualized for complex diseases. This work is aimed to improve the understanding of the concept of heritability, i.e. the proportion of disease variability explained by genetic factors, by showing how the AF and heritability can be formally related.

In summary, we wanted to:

- ◇ Develop a package in the statistical software R [27] that incorporates the estimating strategies for the AF used in the most common epidemiological study designs.
- ◇ Develop a method for estimating absolute risks and the AF for clustered data with time-to-event outcomes and implement this method in the R package.
- ◇ Develop a method to estimate the AF in IV analysis and implement the method in the R package.
- ◇ Create a better understanding of the AF with overall genetic risk as the exposure for complex diseases.
- ◇ Show how the AF and heritability are formally related.

In order to put this work in the greater context, the models used for confounding adjustment and estimation strategies for estimating the AF are described in the following sections.

3 Model-based estimation of the AF in classical sampling designs

Study design plays an important role in epidemiology and has consequences for the statistical analysis and AF estimation. In a cross-sectional study design, the sample is a random selection of subjects of the populations at a specific time point. However, for rare diseases, such sampling may contain few or no cases. In order to ensure that there is enough cases for statistical analysis, and improve statistical power, a case-control study design is often used [5].

In a case-control study design, cases are first selected and controls are then sampled from the same population as the cases. Such sampling scheme implies that the outcome prevalence is fixed by study design and that the estimate of the outcome prevalence in the sample will not be representative of the outcome prevalence in the population.

A limitation with both cross-sectional and case-control study designs is that they do not contain information on the time between exposure and disease, which is important for understanding the disease etiology [5]. The cohort study design is used to retrieve information on the time dimension of the disease development by sampling the study participants from the population at time 0, and follow the subjects until they experience the event, censoring or end of study. Recently, the AF function, seen in Eq. (1.4), was developed for cohort study designs.

Depending on the study design, different statistical methods are used to estimate the AF. In the following sections we give an overview of the statistical methods and the estimation procedures used for estimating the AF.

3.1 Estimation of the AF based on logistic regression

The AF, as defined in Eq. (1.2), contains two elements: the counterfactual probability $\Pr(Y_0 = 1)$ and the factual outcome prevalence $\Pr(Y = 1)$. While $\Pr(Y = 1)$ usually can be estimated directly from the data, we need model-based adjustment in order to estimate the counterfactual $\Pr(Y_0 = 1)$ in observational data. With 'model-based adjustment' we mean that we model the probability of the outcome with the confounders as covariates in the model. Since linear or log-linear models may fail to yield probabilities between 0 and 1, the logistic model is the standard model for AF estimation [28]. The logistic regression model can be defined as in Eq. (3.1),

$$\text{logit}\{\Pr(Y = 1|X, \mathbf{Z})\} = g(X, \mathbf{Z}; \boldsymbol{\beta}), \tag{3.1}$$

where $g(\cdot)$ is an additive function of the variables X and $\mathbf{Z}$ indexed by the parameter vector $\boldsymbol{\beta}$. For example, $g(\cdot)$ could be specified as $\beta_0 + \beta_1 X + \beta_2 \mathbf{Z}$ or contain interactions or other functional forms of X and $\mathbf{Z}$. Logistic regression estimates the parameters $\boldsymbol{\beta}$, which are the log ORs.

In cross-sectional sampling design we use the logistic regression model in Eq. (3.1) to predict the counterfactual outcome prevalence, $\Pr(Y_0 = 1)$, with exposure-level fixed at 0, for each subject. Given that $\mathbf{Z}$ contains all confounders sufficient for confounding control, the average of the predictions, over the sampling distribution of $\mathbf{Z}$, $E_{\mathbf{Z}}\{\Pr(Y = 1 \mid X = 0, \mathbf{Z})\}$ will equal the counterfactual outcome prevalence, $\Pr(Y_0 = 1)$ [9, 12].

In a case-control study design, we cannot estimate absolute probabilities, i.e. $\Pr(Y = 1)$ and $\Pr(Y_0 = 1)$, since the outcome prevalence is fixed by the study design. An alternative is to use Bayes' rule to reformulate the AF, as defined in Eq. (1.2), so that the AF is a function of the adjusted RR [13],

$$\text{AF} = 1 - E_{X,\mathbf{Z}|Y=1}\{\text{RR}(\mathbf{Z})^{-X} | Y = 1\}. \tag{3.2}$$

The expected value is taken over the conditional distribution of X and $\mathbf{Z}$ among the cases and $\text{RR}(\mathbf{Z})$ is the conditional RR defined as in Eq. (3.3),

$$\text{RR}(\mathbf{Z}) = \frac{\Pr(Y = 1 \mid X = 1, \mathbf{Z})}{\Pr(Y = 1 \mid X = 0, \mathbf{Z})}. \tag{3.3}$$

If the disease is rare, the RR can be approximated by the OR. Thus, we may replace $\text{RR}(\mathbf{Z})$ in Eq. (3.2) with $\text{OR}(\mathbf{Z})$, defined in Eq. (3.4),

$$\text{OR}(\mathbf{Z}) = \frac{\Pr(Y = 1 \mid X = 1, \mathbf{Z}) / \Pr(Y = 0 \mid X = 1, \mathbf{Z})}{\Pr(Y = 1 \mid X = 0, \mathbf{Z}) / \Pr(Y = 0 \mid X = 0, \mathbf{Z})}. \tag{3.4}$$

Since the case-control design is motivated by a rare disease, the most common way to estimate the AF from case-control studies is by logistic regression, using the formulation of the AF in Eq. (3.2) [13].

3.2 Estimation of the AF based on the Cox PH model

In cohort studies, we are interested in capturing the time-specific AF by estimating the AF as a function of time t. The AF function in Eq. (1.4) is defined in terms of, $\Pr(T \leq t)$, and $\Pr(T_0 \leq t)$, i.e. the probability of an event at or before time t and the corresponding counterfactual quantity, respectively. Since $\Pr(T \leq t) = 1 - \text{S}(t)$, where $\text{S}(t)$ is the survival function, the estimation of the AF function in Eq. (1.4) is based on modelling the factual, $\mathrm{S}(t)$, and counterfactual, $\mathrm{S}_0(t)$, survival functions [15, 16].

$$\mathrm{AF}(t) = 1 - \frac{\{1 - \mathrm{S}_0(t)\}}{\{1 - \mathrm{S}(t)\}}. \tag{3.5}$$

The survival function is the probability that an individual survive beyond time t and it is defined as

$$\mathrm{S}(t) = \Pr(T > t) = \int_t^\infty f(t)dt, \tag{3.6}$$

where $f(t)$ is some density function. The survival function can also be expressed as in Eq. (3.7),

$$\mathrm{S}(t) = \mathrm{E}_{X,\mathbf{Z}}\{\mathrm{S}(t \mid X, \mathbf{Z})\}. \tag{3.7}$$

where $\mathrm{E}_{X,\mathbf{Z}}\{\mathrm{S}(t \mid X, \mathbf{Z})\}$ is the expected value of the survival function conditional on $\mathbf{Z}$ and X, taken over the distribution of $\mathbf{Z}$ and X.

If $\mathbf{Z}$ contains all confounders sufficient for confounding control at baseline, the counterfactual survival function, $\mathrm{S}_0(t)$ equals the expectation, over the distribution of $\mathbf{Z}$, with $X = 0$ for all subjects, as seen in Eq. (3.8),

$$\mathrm{S}_0(t) = \mathrm{E}_{\mathbf{Z}}\{\mathrm{S}(t \mid X = 0, \mathbf{Z})\}. \tag{3.8}$$

Chen et al. (2010) [15] and and Sjölander and Vandsteelandt (2014) [16] have suggested using the Cox PH model to estimate the factual and counterfactual survival functions in Eq. (3.7) and (3.8) and the AF function in Eq. (3.5).

The Cox PH model is defined in Eq. (3.9),

$$\lambda(t \mid X, \mathbf{Z}) = \lambda_0(t)e^{g(X,\mathbf{Z};\boldsymbol{\beta})}, \tag{3.9}$$

where $\lambda_0(t)$ is the baseline hazard at time t and $g(X, \mathbf{Z}; \boldsymbol{\beta})$ is some function of the exposure X and confounders $\mathbf{Z}$ indexed by the parameter vector $\boldsymbol{\beta}$. In the Cox PH model, we assume that the covariates are multiplicatively related to the hazard which implies that the baseline hazard, $\lambda_0(t)$, does not need to be specified [30].

Based on the estimated Cox PH model we can estimate the conditional survival function in Eq. (3.10),

$$\hat{\mathrm{S}}(t|X_i, \mathbf{Z}_i) = e^{-e^{g(X_i,\mathbf{Z}_i;\hat{\boldsymbol{\beta}})}\hat{\Lambda}(t)} \tag{3.10}$$

where $\hat{\boldsymbol{\beta}}$ is the estimated coefficients from the Cox PH model and $\hat{\Lambda}(t)$ is the estimated cumulative baseline hazard, defined as $\Lambda(t) = \int_{u=0}^{t} \lambda(u)du$. The cumulative baseline hazard can be estimated semi-parametrically by the Breslow estimator [29].

We obtain an estimate of the survival function, $\mathrm{S}(t)$ in Eq. (3.7, by using the conditional survival function in Eq. (3.10) to predict the survival for each subject and taking the sample average. Given that $\mathbf{Z}$ contains all confounders sufficient for confounding control at baseline, the same procedure, but with X fixed to 0 for everyone in Eq. (3.10), yields an estimate of the counterfactual survival function $\mathrm{S}_0(t)$. The estimates of $\mathrm{S}(t)$ and $\mathrm{S}_0(t)$ are then used to estimate the AF function in Eq. (3.5).

4 Methods for clustered data

Data can be clustered, or correlated, in many different ways but the general feature of clustered data is that subjects within a cluster tends to share features. For example, siblings share genetic and childhood environmental factors to a larger extent than unrelated individuals and classmates share their educational environment to a larger extent than students at other schools. The factors shared within clusters may sometimes be important unmeasured confounders and statistical methods that adjust for the cluster-shared factors are important tools in causal inference [31].

An example of informative cluster is shown in the DAG in Figure 4.1. Let two subjects in the same cluster j be indexed by i and i', respectively. Assume that the unmeasured confounders $\mathbf{U}_j$ are shared within the cluster j. The subjects may have different exposure status X_{ij} and $X_{i'j}$, outcome status Y_{ij} and $Y_{i'j}$ and observed confounder status Z_{ij} and $Z_{i'j}$, respectively. Matched data is also clustered data but in comparison with 'natural' clusters, such as twins or school classes, and matched samples are what $\mathbf{U}$ in Figure 4.1 contains [32]. While $\mathbf{U}$ contains all factors shared within the 'natural' clusters, $\mathbf{U}$ will mainly contain the factors that were matched on in a matched cluster.

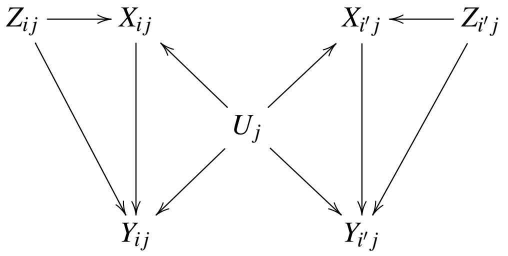

Figure 4.1: Directed Acyclic Graph (DAG) of cluster shared confounders $\mathbf{U}$ for subject i and i' in cluster j with exposure status X, outcome Y and measured confounders Z.

4.1 Clustered data with point-outcomes

For binary outcomes, conditional logistic regression is one way to adjust for factors shared within a cluster. The conditional logistic regression is equivalent to fixed effect regression for binary outcomes and is often used to model matched case-control data [33]. The conditional logistic regression can be defined as in Eq. (4.1)

$$\operatorname{logit}\{\Pr(Y=1|X_{ij}, Z_{ij}, \text{cluster } j)\} = \beta_{0j} + \beta_1 X_{ij} + \beta_2 \mathbf{Z}_{ij}. \tag{4.1}$$

In contrast to the logistic regression in Eq. (3.1), the model in Eq. (4.1) account for the cluster-shared factors by the cluster specific intercept β_{0j}.

An alternative to conditional logistic regression is to adjust for the cluster-shared factors using standard logistic regression with a dummy variable for each cluster [34]. However, when the number of subjects within a cluster is small, and the number of clusters increase, we will in general not get consistent estimates of the parameters in Eq. (4.1) [35]. Conditional logistic regression avoid this problem by adjusting for the cluster-shared factors by conditioning on the cluster rather than estimating the cluster-specific intercepts. This implies that conditional logistic regression cannot be used to estimate absolute probabilities, since β_{0j} in Eq. (4.1) is not estimated.

Thus, for matched case-control data, or other types of clustered data with a rare outcome, we estimate the AF based on Eq. (3.2), where $\mathrm{RR}(Z)$ is approximated by $\mathrm{OR}(Z)$, estimated from a conditional logistic regression model.

4.2 Clustered data with time-to-event outcomes

The stratified Cox PH model in Eq. (4.2) is an analogue to the conditional logistic regression model for time-to-event outcomes.

$$\lambda(t \mid X_{ij}, Z_{ij}, \text{cluster } j) = \lambda_{0j}(t) e^{\beta_W X_{ij} + \gamma Z_{ij}} \tag{4.2}$$

where t is the event-time, X_{ij} and Z_{ij} are the exposure and observed confounders for subject i in cluster j. The stratified Cox PH model is estimated with a conditional likelihood and thus, does not model the cluster-specific baseline hazard $\lambda_{0j}(t)$ in Eq. (4.2). For estimation of the AF based on the Cox PH model, the Breslow estimator was used to estimate the survival function. However, when we have few subjects in each cluster, which is the standard case in family studies, the Breslow estimator will not give a consistent estmate of the cumulative baseline hazards, $\lambda_{0j}(t)$, when the number of clusters grows. Hence, for small clusters, we cannot use the same approach for the stratified Cox PH model as that for the Cox PH to estimate the AF function in Eq. (1.4) [5, 36].

A method often considered for clustered data with point outcomes is the random effect model. The random effect model allows for a cluster-specific intercept, which is assumed to be independent of the covariates in the model and follow some distribution [34]. In contrast to models based on conditioning on the cluster, the random effect model does not adjust for cluster-shared confounders, by the independence assumption, but can be used to estimate absolute probabilities. The frailty model in Eq. (4.3) is the time-to-event analogue to the random effect model

$$\lambda(t \mid X_{ij}, Z_{ij}, \text{cluster } j) = \lambda_0(t) U_j e^{\beta_W X_{ij} + \gamma Z_{ij}} \tag{4.3}$$

where the cluster-specific baseline hazard in the stratified Cox PH model, Eq. (4.2), is factorized into the baseline hazard $\lambda_0(t)$ and the cluster-specific 'frailty' effect U_j, which we assume follows some distribution [37–39].

Thus, there are different limitations with the stratified Cox PH model and the frailty model. While the stratified Cox PH model cannot be used to estimate absolute survival probabilities for small cluster sizes, the frailty model assumes independence between the cluster-specific frailty effect, U_j, and the covariates in the model, which *a priori* rules out cluster-shared confounding [40].

Brumback et al. (2010) [41] propose the 'between-within' (BW) model as a possible solution to this problem for binary data. The BW model was first described by Mundlak (1978) [42] and later by Neuhaus and Kalbfleisch (1998) [43]. The basic idea with the BW model is to include a within cluster effect in order to adjust for cluster shared unmeasured confounding. Sjölander et al. (2013) [40] have described how the BW model can be used for time-to-event outcomes. The BW frailty model can be defined as

$$\begin{aligned} \lambda(t \mid X_{ij}, Z_{ij}, \text{cluster } j) =& \lambda_0(t) U_j e^{\beta_W X_{ij} + \beta_B \bar{X}_j + \gamma Z_{ij}} \\ =& \lambda_0(t) \underbrace{U_j e^{\beta_B \bar{X}_j}}_{U_j^*} e^{\beta_W X_{ij} + \gamma Z_{ij}} \end{aligned} \tag{4.4}$$

where U_j^* collects both the cluster-specific frailty effect and the cluster average of the exposure, $U_j e^{\beta_B \bar{X}_j}$. This 'trick' enables the BW frailty model to allow for dependence between the cluster-specific component and the covariates in the model and thus, to adjust for cluster-shared confounding. In Eq. (4.4), the between-cluster effect β_B captures the dependence between U_j and the cluster-level exposure $\bar{X}_j$. If the observed Z, and cluster-shared unobserved confounders, are sufficient for confounding control, the within-effect β_W, captures the conditional exposure-outcome effect. Thus, by a simple trick when specifying the model, we can use the frailty model to adjust for cluster-shared confounding.

There are several ways to estimate the frailty model. Sjölander et al. (2013) [40] propose assuming a Weibull baseline hazard and a gamma distributed frailty. Other alternatives are to use a semi-parametric approach with an unspecified baseline-hazard modelled by splines [44, 45].

In **project II** we have developed an estimation strategy for the AF function in Eq. (1.4) based on the BW frailty model.

5 Instrumental variable analysis

So far, we have covered methods for model-based adjustment with observed covariates and possibilities to adjust for cluster-shared unobserved confounding. However, we may still have unmeasured confounding that is not shared within the cluster. A natural extension is thus to consider using instrumental variables (IVs) to avoid the problem of confounding.

The motivation behind IV analysis is that by finding a variable G, which is uncorrelated with the confounders $\mathbf{U}$ but correlated with the exposure X. For example, a variable which is randomized in the population but associated with the exposure. If we find such an IV variable, we can mimic a RCT in observational data, and estimate the causal effect [46]. For a variable G to be a valid IV it should fulfill the three requirements: 1) it should be associated with the exposure X, 2) it should not be associated with any of the confounders of the exposure-outcome relationship, $\mathbf{U}$, conditional on the exposure and 3) its effect on the outcome Y should only be mediated by X [47]. Let G denote the IV, in Figure 5.1, crossed arrows denote arrows that would violated the three IV assumptions,

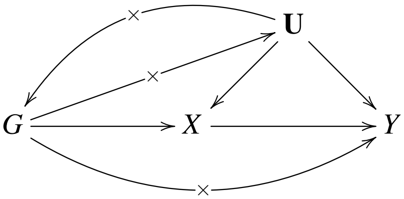

Figure 5.1: *Valid instrumental variable G for the causal effect of X on Y.*

Treatment assignment in an RCTs with non-compliance is often used as an example of an ideal IV to motivate IV analysis. However, in observational data it has been proven difficult to find IVs that simultaneously fulfill the three IV assumptions [48].

An appealing alternative is the Mendelian randomization (MR), where genetic variants, usually single-nucleotide polymorphisms (SNPs), are used as instruments. There are two reasons for the popularity of MR. First, the second IV assumption is likely to hold since genotypes are transmitted from parents to offspring by random assortment, with respect to other genotypes, according to Mendel's first and second law of inheritance [49]. Another reason are the new large sources of genetic data, i.e. UK biobank (UKBB), and results from Genome Wide Association studies (GWAS), that provide researchers with a large set of possible instruments [50].

Since no method for estimating the AF based on IV analysis had previously been developed, we developed an estimator in **project IV**. The most common estimation method for IV analysis is the Two-Stage Least Square (TSLS) regression [51]. The TSLS is, however, limited to continuous exposures and outcomes and for the purpose of estimating the AF we need to use IV estimators for binary outcomes.

The two main methods for IV analysis with binary outcomes are the Two Stage (TS) estimator, which is an analogue to the TSLS for binary outcomes and follows the same principle as the Wald estimator, and the G-estimator, also often denoted as the Structural mean model (SMM) [49]. The aim with both of these estimators is to estimate the causal effect ψ by the structural causal model defined in Eq.(5.1),

$$\eta\{\Pr(Y=1 \mid X,G)\} - \eta\{\Pr(Y_0=1 \mid X,G)\} = \psi X. \tag{5.1}$$

where η is the log link for the causal RR and logit link for the causal OR.

An important limitation with IV analysis are the several, partly untestable, assumptions. Even though procedures for testing IV assumptions 2-3 have been suggested [52, 53], these are not guaranteed to rule out an invalid IV, even asymptotically. Moreover, without additional assumptions it is only possible to estimates bounds for the causal effect in Eq. (5.1) [46, 52]. In order to identify the population causal effect, which we are interested in for AF estimation, we have to assume that the causal effect is constant within levels of the IV, i.e. the ‘no effect modification’ assumption (NEM). The NEM assumption is restrictive and the possibility to test it will depend on the application [54].

The ‘monotonicity’ assumption is an alternative assumption, made in order to identify the local average treatment effect (LATE) [54]. In terms of the RCT with non-compliance, the monotonicity assumption implies that there are no ‘defiers’, i.e. subjects which would take the treatment if not assigned, but refrain from treatment if assigned.

In the following two sections, the two main classes for IV analysis for binary outcomes [49] are described. Both methods relies on the NEM assumption in order to identify the causal effect, ψ.

5.1 The Two Stage estimator

The TS estimator estimate the causal parameter ψ in two steps. In the first step in Eq. (5.2) a model for the exposure-IV association is fitted,

$$h\{\mathrm{E}(X_i \mid G_i)\} = \alpha_0 + \alpha_1 G_i \tag{5.2}$$

where $h(\cdot)$ is some link function. The predictions of X from Eq. (5.2) are used as the dependent variable to fit a model for the outcome Y in Eq. (5.3),

$$\eta\{\mathrm{E}(Y_i \mid \hat{X}_i)\} = \beta_0 + \psi \hat{X}_i \tag{5.3}$$

where $\eta(\cdot)$ is some link function. The TSLS estimator coincides with the TS estimator when linear models are used in both steps. For binary outcomes, a log link in stage 2 is suggested to estimate the causal RR and a logistic link in step 2 for estimating the causal OR [49].

Even though the TS estimator is an appealing analogue to the TSLS estimator for binary outcomes, the TS estimator can only estimate the causal RR consistently for a linear model in stage 1 and a log-linear model in stage 2 [55].

There are thus two important limitations for using the TS estimator to estimate the AF in Eq. (1.2). Firstly, the AF is defined for binary exposures, for which a linear IV-exposure model may not be appropriate. Secondly, and more generally, we cannot use the logistic regression model in stage 2 to obtain a consistent estimate of the causal OR [49, 55].

5.2 The G-estimator

The G-estimator is an alternative to the TS estimator for IV analysis [49]. The G-estimator has been shown to consistently estimate the causal RR [56] and the causal OR [57].

G-estimation is based on the principle that, if the IV assumptions holds, the counterfactual outcome for all subjects unexposed, Y_0, is conditionally independent of the IV, G, given the exposure X, i.e. $Y_0 \perp G \mid X$. Thus, the G-estimator of ψ is the value of ψ which, on average, yields a zero covariance between G and Y_0 [58].

The log causal RR can be estimated by finding the value of ψ that satisfies the equality in Eq. (5.4) [56] by averaging over the sample,

$$0 = \sum_{i=1}^{n} \{G_i - \widehat{E(G)}\} Y_i \exp(-\psi X_i) \tag{5.4}$$

where $\widehat{E(G)}$ denotes the mean of G.

The G-estimation of the causal OR follows a slightly different procedure due to the non-linearity in the logit function, used as link function in Eq. (5.1). This requires an additional estimation of an 'associational model', Eq.(5.5), in order to get a consistent estimate of ψ from the estimating equation in Eq. (5.6) [57].

$$\text{logit}\{\Pr(Y = 1 \mid X, G)\} = m(X, G; \beta), \tag{5.5}$$

where $m(\cdot)$ is some function of X and G.

To minimize the problem of model-misspecification in Eq. (5.5), which may introduce bias, a saturated model can be used if both X and G are categorical with few levels [57].

Vansteelandt et al. (2011) [57] have shown that the estimating equation in Eq. (5.6) yields a consistent estimate of the log causal OR ψ,

$$0 = \sum_{i=1}^{n} \{G_i - \widehat{E(G)}\} \times \text{expit}\{m(X_i, G_i; \hat{\beta}) - \psi X_i\} \tag{5.6}$$

where expit(x)$\equiv \exp(x)/\{1+\exp(x)\}$, $\widehat{E(G)}$ denotes the mean of G_i and $m(X_i, G_i; \hat{\beta})$ is the predictions from the associational model defined in Eq. (5.5).

While the TS estimator is guaranteed convergence, since it is based on two fully parametric models, the G-estimator may not converge. The convergence of the semi-parametric G-estimator relies more heavily on the IV and the other model assumptions. For example, non-convergence is more common when the IV-exposure association is weak, i.e. violation of IV assumption 1 [59].

6 Variance estimation with the sandwich estimator

For some non-standard or nested parameter, i.e. a parameter that is a function of other parameters, such as the AF, the bootstrap and the sandwich estimator (or Robust standard errors and Huber-White sandwich estimator [51, 60]) can be used to estimate the variance.

Both the sandwich and the bootstrap estimator rely on large sample properties [51], but differ in that the bootstrap estimator is based on resampling while the sandwich estimator is based on an analytical expression. The bootstrap may provide an asymptotic refinement, when used carefully, compared to the sandwich estimator [51] but the sandwich estimator has the advantage of being computationally more efficient in large datasets. For this purpose, the sandwich estimator is preferable for software implementations.

In clustered data, the standard errors are typically underestimated but with the sandwich estimator, the within-cluster correlation can easily be accounted for [51]. For these reasons, we have used the sandwich estimator for estimating the variance of the AF in our software `AF`, even though other variance estimators has been proposed for the AF [61, 62].

6.1 Theory for M-estimators

The sandwich estimator is derived from the theory for M-estimators [60]. M-estimators are a broad class of estimators, incorporating non-linear as well as standard maximum likelihood estimators, all based on estimating equations. A detailed description of M-estimators are given in Stefanski and Boos (2002) [60] and a brief description is given here.

Let θ be the parameter of interest, and let $\mathbf{Y}$ be a generic variable where, $\mathbf{Y} = Y_1, \ldots, Y_n$ are independent, but not identically distributed, observations. Moreover, $\hat{\theta}$ is the M-estimator of θ. The M-estimator is the value of $\hat{\theta}$ which satisfies the Eq. (6.1),

$$\sum_{i=1}^{n} \phi(Y_i, \hat{\theta}) = 0 \tag{6.1}$$

where $\phi(\cdot)$ is a known function that does not depend on i or n [60]. Eq. (6.1) is then the estimating equation for θ. From the theory of M-estimators,

$$n^{1/2}(\hat{\theta} - \theta) \overset{p}{\sim} N(0, \Sigma). \tag{6.2}$$

By the Taylor series expansion of Eq. (6.2), we can get an expression of the asymptotic variance $\text{Var}(\hat{\theta})$, shown in Eq. (6.3)

$$\text{Var}(\hat{\theta}) = A(\theta_0)^{-1}B(\theta_0)\{A(\theta_0)^{-1}\}^T \tag{6.3}$$

where θ_0 is the asymptotic solution to the estimating equation in Eq. (6.1). The 'bread', $A(\theta_0)$, is the expected value of the first derivative of the estimating equation $\phi(Y_i, \hat{\theta})$, with the empirical estimator in Eq. (6.4)

$$A_n(\mathbf{Y}, \hat{\theta}) = \frac{1}{n}\sum_{i=1}^{n}\left[\partial\phi(Y_i, \hat{\theta})\right] \tag{6.4}$$

and $B(\theta_0)$ is the variance of the individual contributions from the estimating equations, with the empirical estimator in Eq. (6.5)

$$B_n(\mathbf{Y}, \hat{\theta}) = \frac{1}{n}\sum_{i=1}^{n}\phi(Y_i, \hat{\theta})\phi(Y_i, \hat{\theta})^T. \tag{6.5}$$

In maximum likelihood theory, $B(\theta)$ denotes the variance of the score contributions and $A(\theta)$ is the Fisher information, i.e. the hessian of the score function, since these coincide in ordinary maximum likelihood estimation, Eq. (6.3) simplifies to $B(\theta)$.

Correct variance for clustered data is given by within-cluster versus between-cluster summation, where the clusters are assumed to be independent. In the Appendix of **project II**, this procedure is described.

6.2 Estimation of the variance of the AF

In order to implement the sandwich estimator for the AF, we formulate an estimating equation for each parameter p, used in the estimation of the AF. This results in a system of estimating equations, that takes the dependence between parameters into account. From the system of estimating equations the sandwich estimator in Eq. (6.3) is estimated, resulting in a square $p \times p$-matrix, where p is equal to the number of estimating equations.

Depending on which estimating strategy that was used for estimating the AF, the procedure to extract the variance of the parameter of interest from the sandwich estimator differ. If the AF is estimated as a ratio between a counterfactual and factual quantity, as defined in Eq. (1.2) and Eq. (1.4), the delta method is used to get the variance estimate of the AF. The variances of factual and counterfactual, $\Pr(Y = 1)$ and $\Pr(Y_0 = 1)$, respectively, are extracted from the sandwich matrix. If the AF is estimated based on Eq. (3.2), the variance can directly be retrieved from sandwich estimator. Further details are given in the articles of **project I, II** and **IV**.

7 The AF with a genetic exposure

The AF was initially defined with an environmental exposure in mind, but the exposure of interest could as well be genetic. The most widely used estimation strategy of the AF with a genetic exposures is to define a single, or set of, risk increasing SNPs as the exposure [63, 64].

In the process of detecting the SNPs that contribute to the disease risk, GWAS traditionally uses a stringent significance threshold in order to cope with the multiple testing problem [65]. Consequently, genetic variants with small-effects, in a given sample size, will be neglected. For complex traits, where a large set of SNPs contributes to the overall genetic disease risk [65], this imposes a problem since the SNPs used to estimate the AF will not represent the overall genetic risk. Thus, the approach to estimate the AF based on a single, or multiple, SNPs as exposures is, to this date, difficult to extend to complex traits with the overall genetic risk as the exposure of interest.

Ramakrishnan and Thacker (2012) [66] define an exposed subject as a subject with an affected co-twin and use twin data to estimate the AF for the overall genetic risk. They use the estimation strategy proposed by Levin (1953) [1] in Eq. (7.1),

$$\mathrm{AF} = \frac{P_X(\mathrm{OR}-1)}{1+P_X(\mathrm{OR}-1)} \tag{7.1}$$

where P_X denotes the population exposure prevalence, i.e. the outcome prevalence among co-twins, and the OR is estimated from a logistic regression. Ramakrishnan and Thacker (2012) [66] separate the AF into an AF for the shared environment between twins, AF_c, and an AF for heritability, AF_a. These two AFs are estimated by plugging in the estimate of the OR for the shared environment, $\widehat{\mathrm{OR}}_c$ or the OR for heritability, $\widehat{\mathrm{OR}}_a$, in Eq. (7.1), respectively. The estimates of OR_c and OR_a are derived from the coefficients of the logistic regression in Eq. (7.2),

$$Y = \beta_0 + \beta_1 X + \beta_2 Z + \beta_3 XZ \tag{7.2}$$

where Y is the disease status in the twin and the exposure X is a binary indicator of the disease status in the co-twin and Z is an indicator of zygosity.

Even though Ramakrishnan and Thacker (2012) [66] attempt to capture the overall genetic risk in complex diseases, having an affected co-twin cannot cause the disease. Rather, it is the genetic set-up shared between the twins that is the cause and thus, the exposure of interest. An alternative way to model, and conceptualize, the overall genetic risk is by the liability threshold model.

7.1 The liability threshold model for complex traits

Complex diseases are diseases caused by a large set of environmental and genetic factors, each with a small contribution to the overall disease risk. Thus, the distribution of the disease risk can be well described by an underlying continuous liability scale following a normal distribution [67]. The liability threshold model is depicted in Figure 7.1 and the observed outcome status can then be regarded as an indicator of whether the individual has a liability that exceeds the threshold β on the liability scale [68].

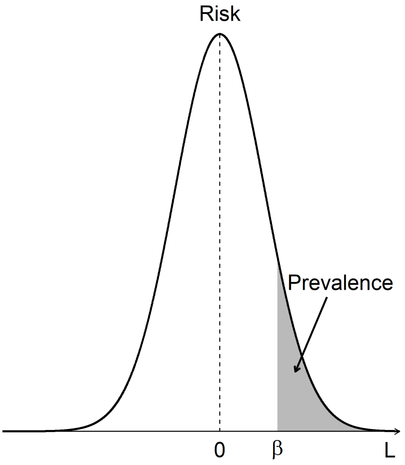

Figure 7.1: *The liability threshold model*

Based on the liability model we may decompose the overall liability into environmental (E) and genetic (G) effects on the liability (L), as in Eq. (7.3).

$$L = G + E. \tag{7.3}$$

With the liability model, we can conceptualize the exposure as the overall genetic risk G and consider some intervention that shift in the genetic risk, G, to lower risk levels. In **project III**, we show how the liability model can be used to formulate the AF with the overall genetic risk as the exposure. In this context, the concept of heritability may help us to quantify the overall genetic risk.

7.2 Heritability

Heritability is a measure of the proportion of variation in a disease that is due to genetic risk defined in Eq. (7.4) [67].

$$h^2 = \frac{\sigma_G^2}{\sigma_L^2} \tag{7.4}$$

where σ_L^2 denotes the overall disease variation and σ_G^2 denotes the genetic variation of the disease. The challenge in estimation of heritability is to separate genetic variation from the overall variation of the disease. For this purpose, heritability is estimated using data on related subjects [69, 70]. One way to estimate heritability is by the ACE model, based on the liability threshold model [70].

The ACE model identifies the disease variation due to genetics by compairing the disease prevalence between monozygotic twins, assumed to share the whole genome, and dizygotic twins, assumed to share around 50% percent of genome [71]. However, the ACE model relies on several strong assumptions of: no assortative mating, no effect of genetic dominance, additive genetic effects, no gene-environment interactions and normally distributed genetic and environmental components. Thus, in practice, estimates of heritability based on the ACE liability model may suffer from bias [69, 72].

The definition of heritability in Eq. (7.4) may look simple, but communicating the meaning of a ratio of two variances is not trivial, which makes heritability difficult to interpret [69]. One common misinterpretation of heritability is that it is the proportion of cases that would have been avoided if no one would have had the harmful genetic composition causing the disease, i.e. the AF with a genetic exposure [69].

Even though both the AF and heritability measure the effect of a genetic factor on a trait in a specific population, the measures estimate different quantities. While heritability estimate the proportion of variance in a disease explained by genetic variation in a population, the AF estimates the proportion of a disease that would be avoided, had a genetic component been eliminated in the population. Nevertheless, both concepts are important in order to understand the disease burden due to genetics. In **project III**, we formalized the relationship between heritability and the AF for complex diseases using the liability threshold model.

8 Summary of projects

8.1 Project I

The use of a statistical method will in practice depend on its availability in statistical software. Even though model-based estimation of the AF has been developed for most statistical software [73], there has been no uniform tool for estimating model-based AF for various study designs in the statistical software `R`. When we started this project in 2015, there were only three packages available at `CRAN`: `epiR` [74], `attribrisk` [75] and `paf` [76].

Each of the available packages had its own limitations. For example, the function `epi.2by2` in the `epiR` package which estimates the AF for cross-sectional, case-control and cohort study designs does not allow for model-based confounder adjustment. Moreover, the `attribrisk` package estimates the AF for case-control and cross-sectional study designs but relies on the 'rare-disease' assumption and is thus restricted to case-control studies in practice. For cohort study designs, the AF can be estimated with the `paf` package using Cox proportional hazard regression for confounder adjustment. The main limitation of the `paf` package is that it does not handle large datasets.

Another limitation is that none of these packages provide accurate standard errors for clustered data [77]. In our package `AF` we aimed at solving the limitations in the other packages and creating a uniform tool for estimating the AF in different study designs.

For example, we made it possible to use large datasets and used the analytical sandwich estimator in order to reduce the computation time. We also made it possible to calculate correct standard errors for clustered data. In the latest version of the package, the AF is estimated based on the logistic regression (`AFlogit`), conditional logistic regression (`AFclogit`) and a Cox PH model (`AFcoxph`). Moreover, our later developments of estimation strategies of the AF in **project II-IV** were implemented in the `AF` package.

8.2 Project II

In **project II**, we developed an AF estimator which enable adjustment for cluster-shared unmeasured confounding. We used the BW frailty model and assumed a Weibull distributed baseline hazard and a gamma distributed frailty effect, as suggested in Sjölander et al. (2013) [40]. We showed how the BW frailty model can be used to estimate the counterfactual survival function $S_x(t)$, if all subject would have been exposed at exposure level x, by standardized survival. Based on these results, we describe an estimation stategy for the AF function defined in Eq. (1.4).

To illustrate our developed methods, we investigated if the association that has been observed between preterm birth and ADHD persists if we adjust for factors shared within sibling pairs [78]. For this purpose, we used Swedish registry data to create a birth cohort followed between 3-18 years of age, or until 2013, whichever came first. In the cohort, we considered children born by the same mother and father as clusters and ended up with 667,282 children from 305,938 families.

The outcome was defined as 'time-to-ADHD diagnosis/ medication' and the exposure was defined as being born before week 37 of gestation. We adjusted for the non-shared potential confounders: birth order, sex, maternal and paternal age at childbearing and maternal smoking during pregnancy.

We compared the stratified Cox PH model, the frailty model and the BW frailty model concerning the estimated effect of preterm birth and ADHD diagnosis. While the hazard ratios estimated in the stratified Cox PH model and the BW model were both around 1.08, the hazard ratio was 1.27 in the ordinary frailty model. The results illustrated our analytical understanding of the relationships between the models, in which the BW frailty model yielded a similar estimate as the stratified Cox PH model. The results showed that adjusting for sibling-shared factors attenuates the effect of preterm birth on ADHD risk, which may be an indication of cluster-shared confounders of the observed association between preterm birth and ADHD in the frailty model.

The estimated AF function based on the BW model ranges between 1% to 0.5% from 3 years of age to 18 years of age (or end of follow-up). Thus, due to the small effect of preterm birth on ADHD, and the low prevalence of preterm birth (9.6%), an intervention which could prevent all preterm births would have an impact of between 0.5-1% on the risk of being diagnosed (or medicated for) ADHD before 18 years of age in the Swedish population.

This project resulted in the function `AFfrailty` implemented in the package `AF`.

8.3 Project III

In order to improve our understanding of the disease burden due to overall genetic risk, we showed how it was possible to conceptualize the AF with an overall genetic exposure using the liability model and heritability in **project III**.

To formulate the AF with an overall genetic exposure, we used the liability model to conceptualize an intervention on the overall genetic risk. We assumed that the liability can be described as a linear combination of genetic and environmental effects, both normally distributed. Furthermore, we consider some hypothetical intervention targeting the genetic risk, such that the genetic risk distribution is shifted to lower levels for all subjects targeted by the intervention. However, since not all subject may benefit from such intervention, i.e. those who already have a low genetic risk, we allowed to model the intervention with a limited target group consisting of those with the highest genetic risk. By using the liability model, we could show how heritability occurs naturally as a parameter in the formulation of the AF. Thus, based on the assumptions similar to those of the ACE model, the AF with an overall genetic exposure can be formalized as a function of intervention effect, disease prevalence, target group size and heritability.

As an illustration of our analytical derivations, we used two real examples. In the first example, we showed how the same intervention of blood pressure and cholesterol lowering medication on stroke and acute myocardial infarction may have different AFs due to the differences in heritability and disease prevalence between the two diseases. In the second example, we turn our focus on how different intervention strategies to prevent breast cancer influence the AF differently for a given heritability and disease prevalence. In both examples, we used estimates of heritability from twin-studies.

In **project III** we concluded that it is possible to formally show how the AF and heritability are related, and formulate the AF for genetic exposures in complex traits. The relationship is, however, highly non-linear and relies on several parameters and assumptions. Since the formula for the AF may be difficult to interpret analytically, we provide the shiny app `afheritability` [79] in the package `AF` to allow the user to investigate the relationships between heritability, prevalence, intervention effect size and target group size graphically.

8.4 Project IV

Unmeasured confounding is a major problem in observational studies. In **project II** we developed a method for confounder robust estimation of the AF for situations in which unmeasured confounding is shared within clusters. However, for situations when clustered data is not available, or we expect that the unmeasured confounding is not shared within clusters, IV analysis is an alternative to handle the problem with unmeasured confounding.

In **project IV**, we developed an AF estimator for IV analysis. We compared the two main IV estimators for binary outcomes: the TS and the G-estimator and showed how the AF can be estimated based on either estimator. The TS and the G-estimator are used to estimate the causal parameter ψ, which may be either the log causal RR or log causal OR.

The first step to estimate the AF was to use the estimated ψ to predict the counterfactual outcome, Y_0, for each individual i. The procedure differed depending on which parameter was estimated, and by which estimator. In the next step, the predictions of Y_{0i} were used to estimate the counterfactual outcome prevalence, $\Pr(Y_0 = 1)$, used to estimate the AF as defined in Eq. (1.2).

Previous studies have observed a reverse association between education level and risk of coronary heart disease (CHD) [80, 81]. A recent MR study used genetic variants as IVs to prove that education has a causal effect on CHD risk [82]. In **project IV**, we used the genetic variants from Tillmann et al. (2017) [82] to illustrate our developed methods and estimate the AF of having a university education on CHD risk with data from the UKBB.

The dataset from the UKBB consists of 267,506 individuals, aged 40-69 years from different study centers across the UK who participated in the UKBB project during 2006-2010. We made two primary analyses: one 'observational', in which we estimated one undjusted OR and one conditional OR, adjusted for the potential confounders: 'maternal smoking during pregnancy', 'comparative body size at age 10', 'age' and 'sex'.

In the 'observational' results, the unadjusted OR was 1.52 with an AF of 23%. When adjusting for the potential confounders, the conditional OR was slightly lower, 1.49, with an AF of 21%.

In the IV analysis, all estimates, independent of choice of parameter or estimation method, were around 1.88-1.99, indicating an almost doubled risk of CHD among those without a university degree. The corresponding AF was around 33-34%. Adjustment for potential IV-outcome confounders 'maternal smoking during pregnancy' and 'comparative body size at age 10', yielded slightly attenuated effect estimates ranging between 1.84-1.93, with an AF of 32-33%. All results were significant on a 5% significance level.

Assuming that the IV assumptions hold, the results would indicate that some confounders, with opposite directions on the exposure and the outcome, confound the crude associations. However, in this application, we have several reasons to expect that the IV assumptions do not hold. We discuss possible violations, solutions and stressed the importance of a critical assessment of the IV assumptions when interpreting the results from IV studies.

This project resulted in the function `AFivglm` in the package `AF`.

9 Discussion

In this work, we have focused on how a causal interpretation of the estimated AF can be obtained by using tools from causal inference and the study design. One essential condition for estimating causal effects in observational data is that of no unmeasured confounding. In practice, unmeasured confounding is expected in observational studies and hampers estimation of causal effects. In **project II** and **IV**, we have shown how clustered data and instrumental variables offer solutions to the problem of unmeasured confounding in some settings. Moreover, by providing a user-friendly R-package for model-based estimation of the AF, we intended to make the methods in causal inference more accessible for epidemiologists and improve estimation of the AF.

With the emergence of larger sources of genetic data, the AF with genetic risk factors has been suggested [63, 64]. In **project III** we show how it is possible to conceptualize the AF with a genetic exposure even if the disease is complex and how this approach also formalize the relationship between heritability and the AF. A result which can be used to improve our understanding of the disease burden of some complex disease due to genetic factors.

The overall aim of this work has been to create a unified framework for AF estimation. The contributions to AF estimation presented in this work is however only a small part of all possible extensions. In the following section we further discuss current limitations and possible directions for future method developments to improve AF estimation in observational studies.

9.1 Limitations and future directions

The methods we proposed to handle the problem with unmeasured confounding suffers from their own limitations. For example, in **project II** we use conditional models to adjust for cluster shared unmeasured confounding. The rationale for conditioning on the cluster is that we expect the unmeasured confounders to be shared within the cluster. It has, however, been shown that in the presence of non-shared confounding and measurement error, the bias will be amplified when we condition on the cluster [83].

Furthermore, in clustered data, we may encounter the problem of ‘carryover effects’, i.e. that subjects within a cluster influence each other’s exposures or outcomes. Carryover effects may yield different types of bias depending on the type of carryover effect [84]. Thus, even though methods that conditions on clusters are promising for adjustment of unmeasured cluster-shared confounding, we have to be aware of multiple sources of bias.

IV analysis has been regarded as another promising tool to identify causal effects in the presence of unmeasured confounding but relies on three assumptions, out of which two, to a large extent, are untestable. Thus, in some sense, IV analysis solves the problem of the untestable ‘no unmeasured confounding’ assumption with other untestable assumptions [46]. Another limitation is the difficulty in finding IVs that simultaneously fulfill the three IV assumptions. MR has been regarded as a solution to this problem, since a large set of SNPs are available for most traits. But with a large number of SNPs as IV (if used as a score), each with a somewhat unknown biological mechanism, justification of the IV based on subject-matter knowledge is often unfeasible [53, 85, 86].

Robust methods have been developed to handle violations of IV assumption 2-3 for some SNPs [87]. Using robust methods could potentially solve the problem of violations of the IV assumptions in some cases, but to this date, all suggested methods have important limitations [87]. Nevertheless, it could be interesting to explore the possibility to incorporate the robust methods with the AF estimation strategy proposed in **project IV**.

Moreover, while the TS estimator relies on distributional assumptions on the IV-exposure relationship, which yield inconsistent estimates of the causal RR and causal OR when the exposure is binary [49], the G-estimator is not guaranteed convergence. Solving either of these limitations would improve IV estimation in situations with binary exposures and outcomes.

Our application in **project IV** did not only highlight the need of care and consideration when using MR analysis, we also recognized the importance of looking into mediation of the exposure-outcome effect. Methods for mediation with AF [20, 21] could be used to improve the design of efficient interventions by creating a better understanding of the population impact of each mediating factor. Moreover, when considering mediation, it would also be interesting to account for interference, i.e. subjects influence on each other’s mediators [88].

Despite the limitations of the methods used in **project II** and **IV**, when used correctly, and with awareness, the methods may, at worst, provide insights into the problem of confounding and at best yield a more confounding robust estimate of the AF.

So far we have assumed that the problem of confounding is known by the researcher, in practice, the crucial problem is that the researcher has to rely on the currently available subject matter knowledge, which, in many cases, may be limited, i.e. the DAG may be wrong. One way to model different confounding scenarios and assess the problem with unmeasured confounding is by sensitivity analysis.

Methods for sensitivity analysis has been developed for the AF [89, 90] and could be used, and extended, in multiple ways. One interesting extension would be to combine the analytical expression for the bias term derived for carryover effects in clustered data [84] with sensitivity analysis for the AF. Other extensions would be to develop the tools for sensitivity analysis for the AF to continuous exposures and in IV analysis.

At a first glance, the topic of **project III** may seem outside the scope of this thesis, but the question regarding the role of genetic exposures is an important issue in epidemiology [91]. A possible application of our results in **project III** would be to use the recently developed genome-wide polygenic scores (GPS) [92], developed for various complex diseases, to consider an intervention which move those at the highest genetic risk to normal risk levels. Translating the results in studies such as Khera et al. (2018) [92] into AF estimates would be a powerful tool to communicate the disease burden for a large set of complex diseases.

Throughout the majority this work, we have considered the AF defined for a binary exposure. However, in many situations it may not be realistic, or of interest, to investigate a complete elimination of the exposure. For example, no realistic intervention to this date can eliminate all cases of preterm birth, as discussed in **project II**. A more general definition of the AF, that allows for continuous exposures, is the generalized impact fraction (GIF) [22–24].

By implementing the GIF as an alternative in in the `AF` package, we would allow the user to model more realistic intervention scenarios. Moreover, many intervention strategies do not only target a single, but several, potential risk factors. For example, interventions to reduce CHD risk by promotion of a healthy lifestyle. Hence, when considering more realistic ways to model the AF, it would also be useful to allow for multiple exposures [93].

In **projects I** and **II** we discuss, and develop, the AF as a function of time. We assume that the exposure and confounders occur at baseline, but in most applications the exposure, and confounders, may be time-varying. Methods for AF estimation that allows for time-varying exposures and confounders, as well as competing risks and multiple states, have been developed [25, 26]. So far, these methods have not been implemented in any R-package and could potentially be part of our package `AF`.

10 Conclusion

The problems discussed in this work mainly relates to the estimation of the AF but are shared by the general field of causal inference. Essentially: how can we obtain an estimate of a causal effect when using observational data? This well-debated topic in epidemiology and statistics still challenge researchers [4], and will continue to do so. One reason lies in the challenge to conceptualize what we mean with a *causal* effect, a question that has been debated within philosophy for centuries [8]. Without a clear definition of causality, traditional statistics have previously limited their scope of attention to the estimation of associations. Even though knowledge on the association between various factors may be interesting and important, it is of limited use when designing interventions to improve public health.

The field of causal inference has intended to bridge the gap between the theoretical limitations of causality and the interest in estimating causal effects in epidemiology by creating a conceptual framework for causal inference [8]. This framework is limited to the focus on conditions to identify causal parameters. Thus, causal inference, by itself, is not enough to *prove* causality. Triangulation of evidence, i.e. consideration of different data sources and study designs, plays a central role in this process. Nevertheless, the tools of causal inference are helpful in the process to disentangle the reasons for *why* the available evidence is not sufficient to prove causality. An understanding of the current limitations in a specific study may aid in improving statistical methods, study design and data sources for future studies.

11 Acknowledgements

I started my journey at MEB in the beginning of 2015, thanks to my excellent supervisor **Arvid Sjölander**. I am so greatful to have had you as my supervisor. You have inspired me and challenged me to do things I never thought I would be able to do. I also want to thank my co-supervisor **Yudi Pawitan** for teaching me the likelihood course and **Patrik KE Magnusson**, for all your help with genetics.

Thank you **Zoltán Kutalik** for your generosity in accepting me as a visiting student in your group and **Sina Rüeger** for interesting discussions!

Pre-dissertation committé: **Tyra Lagerberg, Yiqiang Zhan** and **Zheng Ning**. You are not only very smart, but also very kind to take on this mission.

But the journey at MEB would never had been so nice and smooth without **Marie Jansson, Camilla Ahlqvist, Alessandra Nanni** and **Lina Werner**. Your skills, calmness and sensibility, was very comforting in many stressful situations.

At my first year at MEB, I used to stay late at work with **Andreas Karlsson** and **Johan Zetterqvist**. At these occasions, I got to know about magic things such as Emacs and the problem of non-collapsibility. I feel very happy for this introduction to PhD-life and the passion you shared with me for statistics and programming.

The PubMeb with **Andreas Jangmo, Henrik Olsson, Robert Karlsson** and **Johanna Holm** (among others) during my first year was like being catapult into the social heart of MEB, thank you all for these occasions! It was at the pub where I met some of the future Christmas party co-organizers. Thank you so much **Kathleen (Kat) Bokenberger, Shuyang Yao, Cecilia Radkiewicz, Laura Ghirardi** and **Jie Song** for taking part of organizing the Christmas party. Kat, your work was so impressive and you made it all unforgettable! And **Zheng Ning**, not only do I enjoy to discuss everything from driving license to statistical genetics with you, your role as Snape at the Christmas party was epic!

Hannah Bower, you did a fantastic job at the Christmas party but foremost I want to thank you for a good cooperation during the two years when we organized the student seminars. It was a blessing to work with you!

Emilio Ugalde Morales and **Dylan Williams**, thank you for all nice social activities, great company and friendship.

For those who do not know it, R-ladies Stockholm started at MEB! Thank you **Maya Alsheh Ali, Sophie Debonneville, Ashley Thompson, Tyra Lagerberg** and **Aminata Ndiaye** for good discussions and for making this happen.

The lunch company at MEB is also of top quality. I got to know so much about science and life at these occasions. Thank you **Anna Plym** for your Swiss-skills and giving me currage, **Camilla Sjörs, Nelson Ndegwa Gichora, Malin Eriksson, Ida Karlsson, Julien Bryois, Kelli Lehto, Yi Lu, Nghia Vu, Isabella Ekheden, Nurgul Batyrbekova,**

Linda Abrahamsson, Fei Yang, Shuang Hao, Marie Reilly, Wenjiang Deng, Rose Kerubo Otiso Bosire and **Anastasia Lam** (also for proof reading) and all other nice people, sharing your stories and food recipes over lunch.

The running club with **Anders Forss** and **Ulrika Zagai** and the Christmas choir led by **Anna Johansson** and **Gunilla Sonnebring** are other great initiatives at MEB, thank you for your effort to do this!

The corridors of MEB must have something special. It was here I had some of the most enjoyable and educational discussions, high and low. Thank you: **Peter Ström, Xingwrong Liu, Henrik Winell, Alex Ploner, Mark Clements, Patrick Sullivan, Ralf Kuja-Halkola, Paul Dickman, Erin Gabriel, Paul Lambert, Cecilia Lundholm, Pär Sparren** and **Keith Humpreys** for taking your time to talk and share ideas.

Thank you: **Caroline Weibull**, for all great discussions and inspiration, **Tong Gong**, for your great smile and energy, **Therese Andersson** for spreading sensibility and support, **Alessandra Grotta** for your wisdom and sweetness, **Gabriel Isheden** -my life advisor, for your company, care and support and **Tor-Arne Hegvik** for spreading joy and taking care of Gabriel: you are blessings to any workplace!

I got the opportunity to share office with some very kind and social persons (how did I got anything done?). Thank you **Bénédicte Delcoigne, Myeongjee Lee, Rickard Strandberg, Thorgerdur (Tobba) Palsdottir** and **Andreas Karlsson** for sharing, discussing and supporting! Unfortunately, you all moved away (was it me?) but I got some new great office mates. Thank you **Frida Lundberg, Nikolaos Skourlis** and **Pablo Gonzalez Ginestet** for fun discussions.

I would never have started this PhD project without some people. From my time as a statistics student at Uppsala University I want to thank **Adam Taube** for all your support and informal introduction to biostatistics. You really inspired me! Thanks to you who went from classmates to friends (among others): **Shikun Lee**, my R and travel companion, **Oscar Oelrich**, for good advice and great pen-friend skills and **Gustav Sunqvist**, for helping me to see the funny side of it all.

But I would not have continued with statistics if it wasn't for the endless support from **Kalle** and the whole **Landegren** family. I am very grateful for the years I spent with all of you. The memories from the dinners, long forest walks and discussions on science and life will forever be with me.

My dear friends: I am so greateful to have you around!

My parents and Maja: thank you for all your help.

Mormor, thank you for all the support you gave me. Your care for me, our fika moments during my years at Uni and your common sense has, and will always, be with me.

And my **Adrien**, thank you for giving me work-life balance, adventures, interesting discussions and love. I am so fortunate.

Bibliography

[1] M. L. Levin. The occurrence of lung cancer in man. *Acta - Unio Internationalis Contra Cancrum*, 9(3):531–541, 1953.

[2] J. P. Vandenbroucke, A. Broadbent, and N. Pearce. Causality and causal inference in epidemiology: the need for a pluralistic approach. *International Journal of Epidemiology*, 45(6):1776–1786, 2016.

[3] R. M. Lucas and A. J. McMichael. Association or causation: evaluating links between "environment and disease". *Bulletin of the World Health Organization*, 83(10):792–795, October 2005.

[4] S. Greenland. Randomization, statistics, and causal inference. *Epidemiology (Cambridge, Mass.)*, 1(6):421–429, November 1990.

[5] K. J. Rothman, S. Greenland, and T. L. Lash. *Modern Epidemiology*. Lippincott Williams & Wilkins, Philadelphia, US, 2008.

[6] A. Sjölander. *Causal inference in epidemiological research*. Institutionen för medicinsk epidemiologi och biostatistik / Department of Medical Epidemiology and Biostatistics, 2009.

[7] B. MacMahon and T. F. Pugh. *Epidemiology: Principles and Methods*. Little, Brown and Company, Boston, 1970.

[8] J. Pearl. *Causality : models, reasoning, and inference*. Cambridge University Press, Cambridge, 2009.

[9] A. Sjölander and S. Vansteelandt. Doubly robust estimation of attributable fractions. *Biostatistics (Oxford, England)*, 12(1):112–121, January 2011.

[10] O. Gefeller. Comparison of adjusted attributable risk estimators. *Statistics in Medicine*, 11(16):2083–2091, 1992.

[11] J. Benichou. A review of adjusted estimators of attributable risk. *Statistical Methods in Medical Research*, 10(3):195–216, June 2001.

[12] F. Sturmans, P. G. Mulder, and H. A. Valkenburg. Estimation of the possible effect of interventive measures in the area of ischemic heart diseases by the attributable risk percentage. *American Journal of Epidemiology*, 105(3):281–289, March 1977.

[13] P. Bruzzi, S. B. Green, D. P. Byar, L. A. Brinton, and C. Schairer. Estimating the Population Attributable Risk for Multiple Risk Factors Using Case-Control Data. *American Journal of Epidemiology*, 122(5):904–914, January 1985.

[14] Y. Q. Chen, C. Hu, and Y. Wang. Attributable risk function in the proportional hazards model for censored time-to-event. *Biostatistics*, 7(4):515–529, October 2006.

[15] L. Chen, D. Y. Lin, and D. Zeng. Attributable fraction functions for censored event times. *Biometrika*, 97(3):713–726, January 2010.

[16] A. Sjölander and S. Vansteelandt. Doubly robust estimation of attributable fractions in survival analysis. *Statistical Methods in Medical Research*, December 2014.

[17] S. O. Samuelsen and G. E. Eide. Attributable fractions with survival data. *Statistics in Medicine*, 27(9):1447–1467, April 2008.

[18] G. E. Eide and I. Heuch. Attributable fractions: fundamental concepts and their visualization. *Statistical Methods in Medical Research*, 10(3):159–193, June 2001.

[19] C. Rämsch, A. B. Pfahlberg, and O. Gefeller. Point and interval estimation of partial attributable risks from case-control data using the R-package ’pARccs’. *Computer Methods and Programs in Biomedicine*, 94(1):88–95, April 2009.

[20] G. E. Eide and O. Gefeller. Sequential and average attributable fractions as aids in the selection of preventive strategies. *Journal of Clinical Epidemiology*, 48(5):645–655, May 1995.

[21] A. Sjölander. Mediation Analysis with Attributable Fractions. *Epidemiologic Methods*, 7(1), 2018.

[22] H. Morgenstern and E. S. Bursic. A method for using epidemiologic data to estimate the potential impact of an intervention on the health status of a target population. *Journal of Community Health*, 7(4):292–309, 1982.

[23] K. Drescher and H. Becher. Estimating the Generalized Impact Fraction from Case-Control Data. *Biometrics*, 53(3):1170–1176, September 1997.

[24] M. Taguri, Y. Matsuyama, Y. Ohashi, A. Harada, and H. Ueshima. Doubly robust estimation of the generalized impact fraction. *Biostatistics*, 13(3):455–467, January 2012.

[25] W. Zhao, Y. Q. Chen, and L. Hsu. On estimation of time-dependent attributable fraction from population-based case-control studies. *Biometrics*, 73(3):866–875, 2017.

[26] M. von Cube, M. Schumacher, and M. Wolkewitz. Causal inference with multi-state models - estimands and estimators of the population-attributable fraction. *bioRxiv*, March 2019. URL `https://arxiv.org/abs/1903.10315v1`.

[27] R Development Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2008. URL `http://www.R-project.org`.

[28] S. Greenland and K. Drescher. Maximum Likelihood Estimation of the Attributable Fraction from Logistic Models. *Biometrics*, 49(3):865–872, September 1993.

[29] J. P. Klein and M. L. Moeschberger. *Survival Analysis: Techniques for Censored and Truncated Data*. Statistics for Biology and Health. Springer-Verlag, New York, 2 edition, 2003.

[30] D. R. Cox. Regression Models and Life-Tables. *Journal of the Royal Statistical Society. Series B (Methodological)*, 34(2):187–220, 1972.

[31] J. Zetterqvist. *Statistical methods for twin and sibling designs*. Inst för medicinsk epidemiologi och biostatistik / Dept of Medical Epidemiology and Biostatistics, April 2017.

[32] A. Sjölander and J. Zetterqvist. Confounders, Mediators, or Colliders: What Types of Shared Covariates Does a Sibling Comparison Design Control For? *Epidemiology (Cambridge, Mass.)*, 28(4):540–547, 2017.

[33] D. W. Hosmer, S. Lemeshow, and R. X. Sturdivant. *Applied logistic regression*. Wiley, Hoboken, N.J., 2013.

[34] P. D. Allison. *Fixed effects regression models*. SAGE, Los Angeles, 2009.

[35] J. Neyman and E. L. Scott. Consistent Estimates Based on Partially Consistent Observations. *Econometrica*, 16(1):1–32, 1948.

[36] A. Sjölander. Regression standardization with the R package stdReg. *European Journal of Epidemiology*, 31(6):563–574, June 2016.

[37] O. O. Aalen. Effects of frailty in survival analysis. *Statistical Methods in Medical Research*, 3(3):227–243, 1994.

[38] P. Hougaard. Frailty models for survival data. *Lifetime Data Analysis*, 1(3):255–273, 1995.

[39] O. O. Aalen, M. Valberg, T. Grotmol, and S. Tretli. Understanding variation in disease risk: the elusive concept of frailty. *International Journal of Epidemiology*, 44 (4):1408–1421, August 2015.

[40] A. Sjölander, P. Lichtenstein, H. Larsson, and Y. Pawitan. Between-within models for survival analysis. *Statistics in Medicine*, 32(18):3067–3076, August 2013.

[41] B. A. Brumback, A. B. Dailey, L. C. Brumback, M. D. Livingston, and Z. He. Adjusting for confounding by cluster using generalized linear mixed models. *Statistics & Probability Letters*, 80(21–22):1650–1654, November 2010.

[42] Y. Mundlak. On the Pooling of Time Series and Cross Section Data. *Econometrica*, 46(1):69–85, 1978.

[43] J. M. Neuhaus and J. D. Kalbfleisch. Between- and Within-Cluster Covariate Effects in the Analysis of Clustered Data. *Biometrics*, 54(2):638–645, June 1998.

[44] P. Royston and P. C. Lambert. *Flexible Parametric Survival Analysis Using Stata: Beyond the Cox Model*. Stata Press, College Station, Texas, U.S, first edition, 2011.

[45] M. Clements, X.-R. Liu, P. Lambert, L. H. Jakobsen, A. Gasparini, G. Smyth, P. Alken, S. Wood, and R. Ulerich. rstpm2: Generalized Survival Models, January 2019. URL `https://CRAN.R-project.org/package=rstpm2`.

[46] M. A. Hernán and J. M. Robins. Instruments for causal inference: an epidemiologist's dream? *Epidemiology (Cambridge, Mass.)*, 17(4):360–372, July 2006.

[47] G. W. Imbens and J. D. Angrist. Identification and Estimation of Local Average Treatment Effects. *Econometrica*, 62(2):467–475, 1994.

[48] E. P. Martens, W. R. Pestman, A. de Boer, S. V. Belitser, and O. H. Klungel. Instrumental variables: application and limitations. *Epidemiology (Cambridge, Mass.)*, 17(3):260–267, May 2006.

[49] T. M. Palmer, J. A. C. Sterne, R. M. Harbord, D. A. Lawlor, N. A. Sheehan, S. Meng, R. Granell, G. D. Smith, and V. Didelez. Instrumental variable estimation of causal risk ratios and causal odds ratios in Mendelian randomization analyses. *American Journal of Epidemiology*, 173(12):1392–1403, June 2011.

[50] J. Zheng, D. Baird, M.-C. Borges, J. Bowden, G. Hemani, P. Haycock, D. M. Evans, and G. D. Smith. Recent Developments in Mendelian Randomization Studies. *Current Epidemiology Reports*, 4(4):330–345, 2017.

[51] J. M. Wooldridge. *Econometric Analysis of Cross Section and Panel Data*. Mit Press, 2010.

[52] A. Balke and J. Pearl. Bounds on Treatment Effects from Studies with Imperfect Compliance. *Journal of the American Statistical Association*, 92(439):171–1176, September 1997.

[53] M. M. Glymour, E. J. Tchetgen Tchetgen, and J. M. Robins. Credible Mendelian randomization studies: approaches for evaluating the instrumental variable assumptions. *American Journal of Epidemiology*, 175(4):332–339, February 2012.

[54] P. S. Clarke and F. Windmeijer. Identification of causal effects on binary outcomes using structural mean models. *Biostatistics (Oxford, England)*, 11(4):756–770, October 2010.

[55] V. Didelez, S. Meng, and N. A. Sheehan. Assumptions of IV Methods for Observational Epidemiology. *Statistical Science*, 25(1):22–40, February 2010.

[56] J. Bowden and S. Vansteelandt. Mendelian randomization analysis of case-control data using structural mean models. *Statistics in Medicine*, 30(6):678–694, March 2011.

[57] S. Vansteelandt, J. Bowden, M. Babanezhad, and E. Goetghebeur. On Instrumental Variables Estimation of Causal Odds Ratios. *Statistical Science*, 26(3):403–422, August 2011. arXiv: 1201.2487.

[58] O. Dukes and S. Vansteelandt. A Note on G-Estimation of Causal Risk Ratios. *American Journal of Epidemiology*, 187(5):1079–1084, May 2018.

[59] S. Burgess, R. Granell, T. M. Palmer, J. A. C. Sterne, and V. Didelez. Lack of Identification in Semiparametric Instrumental Variable Models With Binary Outcomes. *American Journal of Epidemiology*, 180(1):111–119, July 2014.

[60] L. A. Stefanski and D. D. Boos. The Calculus of M-Estimation. *The American Statistician*, 56(1):29–38, 2002.

[61] S. Greenland. Variance estimators for attributable fraction estimates consistent in both large strata and sparse data. *Statistics in Medicine*, 6(6):701–708, September 1987.

[62] J. Benichou and M. H. Gail. Variance Calculations and Confidence Intervals for Estimates of the Attributable Risk Based on Logistic Models. *Biometrics*, 46(4):991–1003, December 1990.

[63] J. S. Witte, P. M. Visscher, and N. R. Wray. The contribution of genetic variants to disease depends on the ruler. *Nature Reviews Genetics*, 15(11):765–776, November 2014.

[64] T. Wang, H. D. Hosgood, Q. Lan, and X. Xue. The Relationship Between Population Attributable Fraction and Heritability in Genetic Studies. *Frontiers in Genetics*, 9:352, 2018.

[65] H. Shi, G. Kichaev, and B. Pasaniuc. Contrasting the Genetic Architecture of 30 Complex Traits from Summary Association Data. *The American Journal of Human Genetics*, 99(1):139–153, July 2016.

[66] V. Ramakrishnan and L. R. Thacker. Population attributable fraction as a measure of heritability in dichotomous twin data. *Communications in statistics: Simulation and computation*, 41(3), 2012.

[67] D. S. Falconer. The inheritance of liability to certain diseases, estimated from the incidence among relatives. *Annals of Human Genetics*, 29(1):51–76, August 1965.

[68] D. S. Falconer and T. F. C. Mackay. *Introduction to Quantitative Genetics*. Longman, fourth edition, 1996.

[69] P. M. Visscher, W. G. Hill, and N. R. Wray. Heritability in the genomics era - concepts and misconceptions. *Nature Reviews Genetics*, 9(4):255–266, April 2008.

[70] A. Tenesa and C. S. Haley. The heritability of human disease: estimation, uses and abuses. *Nature Reviews. Genetics*, 14(2):139–149, February 2013.

[71] H. H. Maes. ACE Model. In *Encyclopedia of Statistics in Behavioral Science*. American Cancer Society, 2005.

[72] P. H. Benchek and N. J. Morris. How meaningful are heritability estimates of liability? *Human Genetics*, 132(12):1351–1360, December 2013.

[73] C. Cox and X. Li. Model-Based Estimation of the Attributable Risk: A Loglinear Approach. *Computational statistics & data analysis*, 56(12):4180–4189, December 2012.

[74] M. S. w. c. f. T. Nunes, C. Heuer, J. Marshall, J. Sanchez, R. Thornton, J. Reiczigel, J. Robison-Cox, P. Sebastiani, P. Solymos, K. Yoshida, G. Jones, and S. P. a. S. Firestone. epiR: Tools for the Analysis of Epidemiological Data, September 2015. URL `https://cran.r-project.org/web/packages/epiR/index.html`.

[75] L. Schenck, E. Atkinson, C. Crowson, and T. Therneau. attribrisk: Population Attributable Risk, November 2014.

[76] L. Chen. paf: Attributable Fraction Function for Censored Survival Data, February 2014. URL `https://cran.r-project.org/web/packages/paf/index.html`.

[77] E. Dahlqwist, J. Zetterqvist, Y. Pawitan, and A. Sjölander. Model-based estimation of the attributable fraction for cross-sectional, case-control and cohort studies using the R package AF. *European Journal of Epidemiology*, 31(6):575–582, 2016.

[78] K. Lindström, F. Lindblad, and A. Hjern. Preterm Birth and Attention-Deficit/Hyperactivity Disorder in Schoolchildren. *Pediatrics*, 127(5):858–865, May 2011.

[79] E. Dahlqwist, P. Magnusson, Y. Pawitan, and A. Sjölander. afheritability: a tool to visualize the relationship between the attributable fraction and the heritability: https://afheritability.shinyapps.io/afheritability/, 2018. URL `https://afheritability.shinyapps.io/afheritability/`. Accessed: 2018-01-19.

[80] E. B. Loucks, S. L. Buka, M. L. Rogers, T. Liu, I. Kawachi, L. D. Kubzansky, L. T. Martin, and S. E. Gilman. Education and Coronary Heart Disease Risk Associations May Be Affected by Early Life Common Prior Causes: A Propensity Matching Analysis. *Annals of Epidemiology*, 22(4):221–232, April 2012.

[81] Y. Kubota, G. Heiss, R. F. MacLehose, N. S. Roetker, and A. R. Folsom. Association of Educational Attainment With Lifetime Risk of Cardiovascular Disease: The Atherosclerosis Risk in Communities Study. *JAMA internal medicine*, 177(8):1165–1172, August 2017.

[82] T. Tillmann, J. Vaucher, A. Okbay, H. Pikhart, A. Peasey, R. Kubinova, A. Pajak, A. Tamosiunas, S. Malyutina, F. P. Hartwig, K. Fischer, G. Veronesi, T. Palmer, J. Bowden, G. Davey Smith, M. Bobak, and M. V. Holmes. Education and coronary heart disease: mendelian randomisation study. *BMJ (Clinical research ed.)*, 358:j3542, August 2017.

[83] T. Frisell, S. öberg, R. Kuja-Halkola, and A. Sjölander. Sibling comparison designs: bias from non-shared confounders and measurement error. *Epidemiology (Cambridge, Mass.)*, 23(5):713–720, September 2012.

[84] A. Sjölander, T. Frisell, R. Kuja-Halkola, S. öberg, and J. Zetterqvist. Carryover Effects in Sibling Comparison Designs. *Epidemiology (Cambridge, Mass.)*, 27(6):852–858, 2016.

[85] J. Little and M. J. Khoury. Mendelian randomisation: a new spin or real progress? *Lancet (London, England)*, 362(9388):930–931, September 2003.

[86] D. I. Swerdlow, K. B. Kuchenbaecker, S. Shah, R. Sofat, M. V. Holmes, J. White, J. S. Mindell, M. Kivimaki, E. J. Brunner, J. C. Whittaker, J. P. Casas, and A. D.

Hingorani. Selecting instruments for Mendelian randomization in the wake of genome-wide association studies. *International Journal of Epidemiology*, 45(5):1600–1616, 2016.

[87] S. Burgess, D. S. Small, and S. G. Thompson. A review of instrumental variable estimators for Mendelian randomization. *Statistical Methods in Medical Research*, 26 (5):2333–2355, October 2017.

[88] T. J. VanderWeele, G. Hong, S. M. Jones, and J. L. Brown. Mediation and spillover effects in group-randomized trials: a case study of the 4rs educational intervention. *Journal of the American Statistical Association*, 108(502):469–482, June 2013.

[89] Y. Chiba. Sensitivity analysis for unmeasured confounding of attributable fraction. *Epidemiology (Cambridge, Mass.)*, 23(1):175–176, January 2012.

[90] Y. Chiba. A Simple Method for Sensitivity Analysis of Unmeasured Confounding. *Journal of Biometrics & Biostatistics*, 3(6), August 2012.

[91] M. J. Khoury, R. Davis, M. Gwinn, M. L. Lindegren, and P. Yoon. Do we need genomic research for the prevention of common diseases with environmental causes? *American Journal of Epidemiology*, 161(9):799–805, May 2005.

[92] A. V. Khera, M. Chaffin, K. G. Aragam, M. E. Haas, C. Roselli, S. H. Choi, P. Natarajan, E. S. Lander, S. A. Lubitz, P. T. Ellinor, and S. Kathiresan. Genome-wide polygenic scores for common diseases identify individuals with risk equivalent to monogenic mutations. *Nature Genetics*, 50(9):1219, September 2018.

[93] H. Lin, H. G. Allore, G. McAvay, M. E. Tinetti, T. M. Gill, C. P. Gross, and T. E. Murphy. A Method for Partitioning the Attributable Fraction of Multiple Time-Dependent Coexisting Risk Factors for an Adverse Health Outcome. *American Journal of Public Health*, 103(1):177–182, January 2013.