Causal discovery
from mixed and missing data
with applications on ADHD datasets

Proefschrift

ter verkrijging van de graad van doctor
aan de Radboud Universiteit Nijmegen,
op gezag van de rector magnificus prof. dr. J.H.J.M. van Krieken,
volgens besluit van het college van decanen
in het openbaar te verdedigen op maandag 30 oktober 2017
om 16:30 uur precies

door

Elena Sokolova

geboren op 18 juli 1986
te Novosibirsk (USSR)
Promotoren:
Prof. dr. Tom Heskes
Prof. dr. Jan Buitelaar

Copromotoren:
Dr. Perry Groot
Dr. Jeffrey Glennon

Manuscriptcommissie:
Prof. dr. Elena Marchiori
Prof. dr. Pieter Hoekstra, Rijksuniversiteit Groningen
Prof. dr. ir. Jan Lemeire, Vrije Universiteit Brussel, België

The research reported in this thesis has been carried out under the auspices of the Dutch Research School for Information and Knowledge Systems (SIKS), and the institute for Computing and Information Science (iCIS) of the Radboud University Nijmegen.

Copyright © 2017 Elena Sergeevna Sokolova

Printed by: Ridderprint
## Contents

1 Introduction ......................................................... 1  
   1.1 Challenges in analysis of medical data . . . . . . . . . . . . . . . . 1  
   1.2 Causal discovery vs other methods . . . . . . . . . . . . . . . . . . 4  
   1.3 Introduction to causal discovery ............................... 6  
   1.4 BCCD . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 9  
   1.5 ADHD and ASD .................................................. 10  
      1.5.1 ADHD .................................................... 10  
      1.5.2 ASD ..................................................... 11  
   1.6 Outline of the thesis ........................................... 13

2 Handling hybrid and missing data in constraint-based causal discovery to study the etiology of ADHD 17  
   2.1 Introduction ..................................................... 17  
   2.2 Background ..................................................... 21  
   2.3 Related study and motivation .................................. 22  
      2.3.1 Structure learning ....................................... 22  
      2.3.2 Undirected graphical models ............................ 24  
   2.4 Proposed method ................................................ 25  
   2.5 Experimental results .......................................... 29  
      2.5.1 Simulation study ......................................... 29  
      2.5.2 ADHD data ............................................... 35  
   2.6 Discussion and conclusions ................................... 40

3 Computing lower and upper bounds on the probability of causal statements 43  
   3.1 Introduction ..................................................... 44  
   3.2 A Method for Computing Causal Reliability Intervals ............ 45  
      3.2.1 Constrained-Based Causal Discovery .................... 45  
      3.2.2 Estimation of Probability Intervals ..................... 48
4 Indirect link between DAT1 genetic variants and striatal brain activation during reward processing

4.1 Introduction

4.2 Materials and Methods

4.2.1 Participants

4.2.2 Experiment description

4.2.3 Data description

4.2.4 Brain data preprocessing

4.2.5 Causal Modeling

4.3 Results

4.4 Discussion

5 Statistical evidence suggests that inattention drives hyperactivity/impulsivity in ADHD.

5.1 Introduction

5.1.1 Problem description

5.1.2 Causal discovery from observational data

5.1.3 Related models and methodologies

5.1.4 Goal

5.2 Materials and Methods

5.2.1 Materials

5.3 Data analysis

5.4 Results

5.5 Discussion

5.5.1 Interpretation

5.5.2 Related literature on ADHD

5.5.3 Assumptions

5.6 Conclusion

5.7 Appendix

5.7.1 Derivation of the LCD pattern

6 A causal and mediation analysis of the comorbidity between ADHD and ASD

6.1 Introduction

5

6
CONTENTS

6.2 Materials and Methods .............................................. 96
   6.2.1 Participants .................................................. 96
   6.2.2 Measures ..................................................... 97
   6.2.3 Data analyses ................................................. 100
6.3 Results .................................................................. 101
6.4 Discussion ............................................................. 104
6.5 Appendix ............................................................... 107
   6.5.1 Causal modeling ............................................... 107
   6.5.2 Results BCCD ................................................... 109
   6.5.3 Mediation analysis ............................................ 109

7 Conclusions .............................................................. 115
8 Summary ................................................................ 119
9 Samenvatting ............................................................ 123
10 Curriculum Vitae ....................................................... 127
11 Acknowledgments ...................................................... 129
12 SIKS dissertations ...................................................... 131

Bibliography ................................................................ 147
Introduction of the digital technologies led to a significant increase in the available data in medical domain. However, analysis of this data is a challenging task. This thesis discusses one of these challenges, the inference of causal statements from observed data, applied to the analysis of the data about two frequent diseases ADHD and ASD. In the introduction we provide a background about the existing algorithms that can infer causal statements and discuss the limitations these algorithms. In the second part of the introduction we talk about the two main diseases considered in this thesis: ADHD and ASD. At the end of the chapter we guide the reader through the outline of the thesis.

1.1 Challenges in analysis of medical data

In recent years, the introduction of digital technologies has resulted in an enormous increase of gathered data in the medical domain. This data includes patient (and familial) history, treatment plans, patient surveys, and results of clinical studies, being a source of valuable information. Proper analysis of this data can potentially improve the treatment of patients, find unknown risk factors of diseases, or detect comorbid disorders if the causal relationships are understood. To infer valid conclusions from observed patterns in the data, advanced machine learning techniques should be used.

The main problem when working with observational data is uncertainty about the interpretation of inferred correlations. As often said in statistics: “Correlation does not imply causation”. A simple example that can support this statement is a dependency between lung cancer and yellow teeth. The percentage of people with yellow teeth among people with lung cancer is higher than in the general population, which makes these two events correlated. However, neither yellow teeth cause lung cancer, nor lung cancer causes yellow teeth.
Chapter 1. Introduction

Figure 1.1: Model that explains correlation between yellow teeth and lung cancer by a common cause, smoking.

The correlation observed between these variables is determined by a common cause, in particular, smoking that makes teeth yellow and increases the risk of cancer. Thus, there can be no direct causal relationship between these variables (Figure 1.1).

The explanation in Figure 1.1 of the correlation between yellow teeth and lung cancer by a common cause smoking is quite obvious. In many cases, however, it is much more difficult to understand the nature of the observed correlation between two variables. For example, some studies have shown a correlation between drinking wine and lower risk of heart disease. A conclusion that was made by many wine producers is that wine drinking can improve health (Figure 1.2a). There are, however, other possible causal models that may explain this correlation. A lower level of drinking wine between people with a heart disease can be a consequence of a doctor’s prescription to follow a healthy lifestyle to reduce the chance of disease complication (Figure 1.2b). Another explanation can be again a common cause (Figure 1.2c). For example, wealthy people can afford themselves a glass of good red wine every evening but they can also afford better health care and more vacations than the general population which can lead to a lower rate of heart disease. In this case it is hard to say which model is correct solely based on background knowledge, however, causal discovery allows one to derive a causal model in such cases as will be shown in Chapter 4.

Another challenge when analyzing medical data is the data quality. Medical data is often incomplete: some patients may not have information regarding one or more relevant variables. Due to different hospital policies, the same
1.1. Challenges in analysis of medical data

Figure 1.2: Examples of possible causal models that offer an explanation for the correlation between wine drinking and a lower rate of heart disease.

Disease can be treated and/or diagnosed differently, which makes comparisons non-trivial since there is often not a single field in the data that is complete for all patients. Additionally, missing data can often result from the drop outs of participants during clinical studies. Work with that type of missing data can be very complex since these drop outs do not necessarily have a random pattern and can be caused for example by the low efficiency of the treatment. As a result, conclusions drawn from that type of data can be biased which needs to be accounted for in any causal model.

Furthermore, most of the algorithms that are designed to work with missing data rely on particular assumptions that do not always hold in reality. Thus, the choice of algorithm and interpretation of the results should be done very carefully. For instance, many algorithms rely on the assumption that data is continuous and has a Gaussian distribution. This assumption often does not hold for real-world data. Most often data is a mixture of discrete and continuous variables. For example, in medical data, age and symptom counts are continuous, while variables like gender and diagnostic status are discrete. A simple ad-hoc solution for this problem is to discretize the data. This may lead, however, to loss of information and incorrect results of analysis. Moreover, a Gaussian distribution is a very strong assumption that is not always applicable. To analyze non-Gaussian data non-parametric models can be used.

This thesis describes approaches to tackle the problems of medical data analysis discussed above. In particular, it focuses on the causal discovery approach that under particular conditions can infer causal relationships between variables from observed data (Pearl, 2000). One part of this work extends existing state of the art algorithms for causal discovery to be able to analyze complex data sets and answer advanced questions about causal relationships. Another part of this thesis demonstrates an application of these methods to real-world medical data in order to study the factors that drive particular brain disorders (attention...
Chapter 1. Introduction

Figure 1.3: Example demonstrating the difference between feature selection methods. (a) The real world dependencies where “Cancer” is a variable of interest (b) Predictors selected by regression analysis to predict “Cancer” marked in yellow (c) Predictors selected by causal analysis to predict “Cancer” marked in yellow.

deficit/hyperactivity disorder (ADHD) and autism spectrum disorder (ASD)).

The rest of this chapter is structured as follows. Section 2 compares causal discovery with other methods for medical data analysis. Section 3 provides background information about causal discovery. Section 4 describes an introduction about ADHD and ASD and challenges in medical research of these disorders. Section 5 provides a summary of the thesis.

1.2 Causal discovery vs other methods

This thesis considers causal discovery as the main method for analysis of medical data. To explain the motivation of choosing this method, a comparison of causal discovery and other popular methods in the medical domain is provided below.

Common methods for the analysis of medical data, besides causal discovery, are correlation analysis, regression analysis, and structural equation modeling (SEM). These techniques have some similarities but are designed to answer different research questions. To demonstrate the main difference between the three methods, we turn to the lung cancer example in Figure 1.3. This figure shows interactions between different variables (genes, smoking, anxiety, and results of an X-ray) that are associated with the presence of lung cancer.

The goal of correlation analysis is to find associations between variables. It suggests that when the value of one variable changes the value of another variable will likely change as well. The limitation of correlation analysis is the lack of a causal interpretation of these associations, since this association does not necessarily mean a direct causal effect of one variable on another. Variables
Causal discovery vs other methods

Correlation can also be caused by a common cause between variables as mentioned earlier. As a result, in a system where many variables have common causes and/or there are long indirect associations between variables, almost all variables in the data set sampled from this system can be correlated with the variables of interest, while only few actually cause these variables directly. Returning to the model in Figure 1.3, depending on the strength of the association, all variables except for bronchitis will be correlated with the variable lung cancer.

The goal of regression is to find a model that predicts the outcome variable. This model estimates the statistical significance of each predictor as well as its size effect. Based on these parameters, the most relevant predictors can be selected. The advantage of the regression analysis in comparison to correlation is that it will select only the most relevant variables to predict the outcome value including only the variables that belong to the so-called Markov blanket. In some cases it can reduce the number of selected variables significantly in comparison to correlation analysis. The limitation of regression is again a lack of causal interpretation of these predictors, since a predictor is not necessarily a direct cause of the outcome variable. Thus, a predictor is a variable that can explain the outcome, but is not necessarily a factor/cause of the outcome.

Figure 1.3 (b) demonstrates that predictor variables selected by the regression analysis are not necessarily a direct cause of the variable of interest. Applying regression to this problem gives a set of predictors that include the direct causes of the outcome (smoking, genes), the effects of the outcome (X-ray), and other causes of the effects (bronchitis), see Figure 1.3 (b). All these variables help to determine cancer. Smoking and genes are direct causes of cancer, and X-ray is an indicator of the symptoms, which makes these variables valuable for the diagnosis. Interestingly, although bronchitis is not directly associated with cancer it is included in the regression model. It happens because bronchitis can be a cause of bad X-ray results, thus information about its presence/absence should be included. Anxiety is not selected by regression, since it is not associated with cancer directly and does not provide any new information for the diagnosis, when information about smoking is available. This example demonstrates that not all predictors selected by the regression analysis are direct causes of the outcome variable. In studies where the main goal is to find the factors that can be manipulated to improve the value of the outcome, the lack of causal interpretation of selected predictors can be a serious issue.

Causal discovery (Aliferis et al., 2010; Pearl, 2000) in contrast to correlation and regression predicts the consequences of an action or intervention (Guyon
et al., 2008; Liu and Motoda, 1998). Causal discovery provides the most probable causal model that generated the data. Based on this model, one can either select all the variables that help to predict the outcome, or only the variables that cause the outcome variable directly (factors). Applying causal discovery to select factors in the example in Figure 1.3, we reconstruct the causal model (Figure 1.3 (c)) and select the direct causes (smoking, genes) of the outcome (cancer).

Causal discovery is also closely related to Structural Equation Modeling (SEM). Typically, SEMs are used in a confirmatory setting, where a limited amount of specific structures are selected manually based on prior knowledge or assumptions and compared against each other by scoring them on the available data. Causal discovery methods are used in a more exploratory setting, where the structure of SEM is learnt automatically from data. These methods assume that there is some SEM underlying the data and then aim to reason about its structure. Under particular conditions, parts of the structure can be derived from conditional (in)dependencies. A key advantage of causal discovery over fitting different SEM structures to the data is that causal discovery automatically incorporates latent variables and implicitly considers all possible models instead of just a few and does not require any prior hypothesis. In this thesis the main focus is on exploratory analysis, rather than confirmatory.

Causal discovery can solve the limitation of standard statistical techniques when the main goal of the study is to learn the factors causing the disorder. Moreover, it allows studying interactions between different domains of the disorder, by proving a probabilistic graph as an output. The graph visualizes the relations between all variables in one model, which is easy to interpret, and allows one to discriminate between direct and indirect relationships between variables. Because of these advantages, causal discovery is now often used in the analysis of biomedical data (Chen et al., 2007; Maathuis et al., 2010; Schmidberger et al., 2011).

1.3 Introduction to causal discovery

The main idea of causal discovery is based on the principle of conditional independence (Pearl, 2000) Saying that two variables A and B are conditionally independent given C, means that if we know C, learning B would not change our belief in A. A common way to model causal relationships between variables is using Bayesian networks, where a Bayesian network is a pair (G, Θ) where G = (X, E) is a Directed Acyclic Graph (DAG) with a set of nodes X representing domain variables and a set of arcs E; θ_{X_i} ⊂ Θ is a set of parameters representing the conditional probability of variable X_i given its parents Pa_i in a graph G. In that case an edge A → B between variables represents a direct causal link.
from $A$ to $B$. This means that $A$ influences the values of $B$, but not the other way around. Each node can be a parent and/or a child (descendent) of another node. If $A \rightarrow B$ then $A$ is a parent of $B$ and $B$ is a child of $A$. When building a Bayesian network to describe a particular probability distribution we rely on the Markov condition, which states that any node in a Bayesian network is conditionally independent of its non-descendents, given its parents.

To illustrate different conditional (in)dependencies we return to Figure 1.3. Let’s consider correlation that is caused by an indirect effect of one variable on another. For instance, anxiety and cancer are dependent: people with anxiety will have cancer more often because these people smoke more often, which causes cancer. Thus, it is possible to predict the risk of cancer based on the presence of anxiety. If it is known, however, that a person is smoking, information about his/her anxiety becomes irrelevant to predict the risk of cancer, since it can be predicted directly from the smoking status. Conditional independencies between variables allow distinguishing between direct and indirect association.

Another interesting case is conditional dependence. In Figure 1.3 bronchitis and lung cancer are two independent events that both can give an effect on X-ray results. If it is known, however, that a person does not have bronchitis, but has bad X-ray results, a belief that this person has lung cancer increases. This is an example of how two variables (bronchitis and lung cancer) become conditionally dependent given variable X-ray, since both bronchitis, and lung cancer are possible explanations of X-ray results. This pattern is called a V-structure and it allows learning the direction of the effect between variables. Thus, by inferring conditional (in)dependencies under some reasonable assumptions (Spirtes et al., 2000) it is possible to infer causal models from data.

There is a variety of methods that can be used to learn the structure of a causal network. A broad description of methods can be found in Daly et al. (2011). In general, the main two approaches are: constraint-based and score-based. The constraint-based approach works with statistical independence tests. A score-based approach uses a scoring metric to measure the data goodness of fit given a particular graph structure and accounts for the complexity of the network.

To illustrate the mechanisms of causal discovery we describe one of the most popular algorithms in the field, PC algorithm (Spirtes et al., 2000). This is a constraint-based algorithm. The main idea of PC is to detect conditional independencies in the data, based on that build a skeleton of the graph and then orient the edges of the graph. In the first step algorithm starts with a fully connected graph and then tests conditional independencies for all pairs of $X_i, X_j \subset X$ increasing the conditional set of variables $Z$ (starting from an empty set). If $X_i, X_j$ are independent given $Z$, then the edges between them is removed.
Chapter 1. Introduction

Figure 1.4: Example demonstrating different steps of PC algorithm. (a) Start with a fully connected graph (b) Infer the skeleton based on conditional independencies (c) Orient the edges based on conditional independencies.

On the second step, the algorithm orients the edges based on the observed conditional independencies. For example if $X_i, X_j$ were independent, but became dependent, when conditioned on variable $Y$ connected to both $X_i, X_j$, then we observe a V-structure, and we can orient the edges between these variables as shown on Figure 1.4.

It is not always possible to infer the exact structure of the DAG, since it is possible that two different DAGs can entail the same conditional independencies. In that case these two DAGs are called equivalent to one another. All DAGs that are equivalent to a graph $\mathcal{G}$ form an equivalence class of a graph $\mathcal{G}$, where all members are indistinguishable in terms of implied independencies. To represent the members of this equivalence class, a different type of structure is used, known as a partially directed acyclic graph (PDAG). Causal discovery usually relies on the following three main assumptions (Spirtes et al., 2000):

- **Causal Markov Condition**: each variable is independent of its non-descendant conditioned on all its direct causes.

- **Faithfulness assumption**: there are no independencies between variables that are not implied by the Causal Markov Condition.

- **Causal sufficiency assumption**: there are no common confounders of the observed variables in $\mathcal{G}$ that are not members of the set.

The first assumption states that the Markov condition holds. The second assumption that the conditional independencies observed in the graph are equiv-
alent to independencies in its probability distribution. The last assumption sug-
gests that there are no latent variables that can cause dependency between vari-
ables in the set. This is a very strong assumption since in practice there are often
variables that are not measured in the medical data sets. For example, not all
genes can be measured that can potentially cause the association between symp-
toms, or not all common environmental factors are known that can influence the
behavior of the patients.

If the causal sufficiency assumption is not applicable then a so-called maxi-
mal ancestral graph (MAG) can be used to represent the dependencies between
only the observed variables. In contrast to DAGs, MAGs can also contain bi-
directed $X \leftrightarrow Y$ arcs (indicating that there is a common confounder) and undi-
rected arcs $X \circlearrowleft Y$. The equivalence class for MAGs is a partial ancestral graph
(PAG).

1.4 BCCD

In this thesis we work with one of the state-of-the-art algorithms in causal dis-
covery, Bayesian Constraint-based Causal Discovery (BCCD). Claassen and Hes-
kes Claassen and Heskes (2012a) showed that BCCD outperforms reference al-
gorithms in the field and it provides an indication of the reliability of the causal
links that makes it easier to interpret the results and compare alternative mod-
els. We here describe the basic idea of the method. The main two steps of BCCD
are the following:

Step 1. Start with a fully connected graph and perform adjacency search, esti-
mating the reliability of causal relations, for example $X \rightarrow Y$. If a causal
relation declares a variable conditionally independent with a reliability
higher than a predefined threshold, delete an edge from the graph be-
tween these variables.

Step 2. Rank all causal relations in decreasing order of reliability and orient
edges in the graph starting from the most reliable relations. If there is a
conflict, pick the causal relation that has a higher reliability.

Based on the score of the causal relations, we can rank these relations and
avoid propagating unreliable decisions giving preference to more confident
ones. This can solve the drawback of a standard constraint-based method that
can end up with an unreliable result. Moreover, using a Bayesian score we get
a reliability measure of the final output, which makes it easier to interpret the
results and compare with other alternative models. The BCCD algorithm does
Chapter 1. Introduction

not rely on the causal sufficiency assumption, thus it can detect latent variables in the model.

The first step of the algorithm requires estimating the reliability of a causal relations \( L : X \rightarrow Y' \) given a data set \( D \), which is done using a Bayesian score:

\[
p(L : X \rightarrow Y'|D) = \frac{\sum_{M \in M(L)} p(D|M)p(M)}{\sum_{M} p(D|M)p(M)}, \tag{1.1}
\]

where \( p(D|M) \) denotes the probability of data \( D \) given structure \( M \), \( p(M) \) represents the prior distribution over structures and \( M(L) \) is the set of structures containing the relation \( L \). This reliability measure (1) gives a conservative estimate of the probability of a causal relation. Claassen and Heskes approximate the probability \( p(D|M) \) by \( p(D|\mathcal{G}) \), the marginal likelihood of the data given graph \( \mathcal{G} \) that has a closed form solution for discrete variables, known as the Bayesian Dirichlet (BD) metric Heckerman et al. (1995). There is also a closed-form solution when all variables have a Gaussian distribution, called the BGe metric Heckerman et al. (1995).

1.5 ADHD and ASD

In this work we apply causal discovery in particular to neuropsychiatric disorders as part of the European EU FP7 Translational Adolescent and Childhood Therapeutic Interventions in Compulsive Syndromes (TACTICS) project. The main disorder discussed in this thesis is attention deficit-hyperactivity disorder (ADHD). Different data sets were analyzed to understand the etiology of this disorder and interaction of different symptoms. ADHD is highly comorbid with autism spectrum disorder (ASD), thus one part of the research described in this thesis investigates the factors that lead to the comorbidity of these disorders. An introduction to ADHD and ASD is provided below.

1.5.1 ADHD

ADHD is a frequent and highly heritable neuropsychiatric disorder, affecting 5-6% of children (APA, 2000). Symptoms persist into adulthood in up to 60% of the childhood cases. ADHD is characterized by two types of symptoms: hyperactivity/impulsivity and inattention, which can occur separately or combined. In pediatric populations, ADHD is about 2-3 times more common in boys than girls (Boyle et al., 2011), but gender balance is rather equal in adult populations. Several studies report that ADHD patients tend to have lower intelligence quotients (IQ) compared to healthy controls but whether this is circumstantial or not is unclear.
Three neurocognitive domains that show deficits in ADHD are executive dysfunction, temporal integration, and reward/motivation problems. This thesis mainly considers the last domain, difficulties in processing rewards. ADHD patients tend to choose small immediate rewards instead of larger delayed rewards, more frequently than healthy controls (Luman et al., 2005). The evidence for altered reward processing is also present on the neural level. Fronto-striatal pathways, including the orbitofrontal cortex, medial prefrontal cortex and the ventral striatum (VS), play a crucial role in reward processing with dopaminergic innervation of these regions being involved in reward expectancy and receipt. Adolescents with ADHD typically exhibit altered neural (fMRI BOLD) responses to reward in some of these regions, including ventrostriatal and orbitofrontal regions (von Rhein et al., 2015).

The genetics of ADHD is complex and several candidate genes have been associated with ADHD in meta-analyses, among which the genes forming part of dopaminergic transmission, in particular the dopamine transporter gene SLC6A3/DAT1. Genetic variation of the DAT1 gene may affect the functioning of the dopamine transporter caused by individual variation in regulating levels of dopamine (by altering dopamine reuptake efficiency). Another candidate gene is nitric oxide synthase NOS1, an enzyme involved in the production of the retrograde transmitter nitric oxide which is also associated with other impulsivity disorders. Multiple other candidate genes exist including those involved in noradrenergic, serotonergic, cholinergic transmission, in addition to those affecting brain structure on an ultrastructure level (Caylak, 2012).

Many studies have been performed that identify candidate genes, possible cognitive problems, and brain areas with altered performance. Building a link between all these factors and effects, however, is a complicated task. Franke et al. (2009) proposed a complete endophenotype model to describe the relationships between genes, brain functioning, behavior, and disease symptoms. In this thesis causal discovery is applied to infer the endophenotype model from the observed data sets to determine causal relationships between symptoms and possible risk factors of ADHD.

1.5.2 ASD

Autism spectrum disorder (ASD) is a general term for a set of complex brain disorders characterized by difficulties with social interaction, verbal and non-verbal communication, cognitive rigidity, and repetitive behavior. Similar to ADHD, it is a highly heritable disorder that manifests early in life and causes severe impairment in cognitive and emotional functioning such that the individual can have difficulty in integrating in daily life. Different studies show that ADHD
and ASD are comorbid disorders, in particular, 22–83% of children with ASD have symptoms that satisfy the criteria for ADHD (Matson et al., 2013; Ronald et al., 2008), and vice versa, 30–65% of children with ADHD have clinically significant symptoms of ASD (Clark et al., 1999; Ronald et al., 2008).

As well as ADHD, ASD is more prevalent in boys than girls. The exact proportion varies according to the cohort studied: ASD is two to three times more common among boys than girls in childhood population cohorts (Kim et al., 2011), while in the clinical population the difference is even higher (four to five times) (Fombonne, 2009). The first signs of ASD usually arise at an early age (usually before three) and are marked by reduced social interaction and eye contact. ASD occurs in 0.9-1.1% of the general population (Baron-Cohen et al., 2009; Blumberg et al., 2013).

Similar to ADHD, ASD is a highly heritable disorder, with classic autism showing a heritability of more than 90% (Freitag, 2007). The genetics of ASD is complex with a number of different factors interacting. However, recent papers have suggested that the heritability is explained by chromosome rearrangement, gene disorders, monogenic mutations, copy number variance, and other multifactorial genetic problems (Berg and Geschwind, 2012; Devlin and Scherer, 2012).

Recent studies have suggested that ADHD and ASD have shared genetic factors, estimating the overlap to be between 50-72% (Lichtenstein et al., 2010). On the other hand, family studies also suggest a presence of specific genetic risk factors for ADHD and ASD, since genetic overlap does not entirely explain the presence of these diseases. Besides genetic factors, environmental factors can increase the risk of both disorders and be coded by epigenetic markers. Some studies have suggested that early environmental factors including certain pre/perinatal insults can be a common factor for ADHD and ASD, for example, advanced parental age, low birth weight, maternal smoking, and stress during pregnancy (D’Onofrio et al., 2014; Gardener et al., 2009, 2011; Mill and Petronis, 2008).

Although many studies have been performed to unravel the mechanisms underlying the comorbidity of ADHD and ASD, the exact causes of their co-occurrence are still poorly understood. Complicating factors including disorder heterogeneity, etiological complexity, involving numerous genetic and environmental factors makes it difficult to develop an overall model that explains the interactions between all factors and symptoms. In this work an attempt was made to build a model that explains these interactions using a causal discovery algorithm to a large sample of observational data.
1.6 Outline of the thesis

This thesis contains two main parts. The research in the first part is mainly focused on extending causal discovery algorithms to handle complex data sets. The second part of the thesis focuses on the application of the causal discovery methods to the analysis of ADHD and ASD.

Chapter 2 considers two challenges in causal discovery that occur very often when working with medical data: a mixture of discrete and continuous variables and a substantial amount of missing values. Currently there are no methods available that can handle both challenges at the same time. In this chapter a new method is presented that can handle these challenges based on the assumption that data is missing at random and that continuous variables obey a non-paranormal distribution. The validity of this approach is demonstrated on simulated data as well as on two real-world data sets from a monetary incentive delay task and a reversal learning task aiming to improve an understanding of the etiology of ADHD. This work has been recently published (Sokolova et al., 2014, 2015a, 2016c).

In Chapter 3 an extension to constraint-based methods for causal discovery is discussed. These methods typically work through forward chaining: given some causal statements, others can be inferred by applying relatively straightforward causal logic such as transitivity and acyclicity. Starting from the premise that the reliabilities for base causal statements can be estimated, a new approach to estimate the reliability of novel statements is proposed. Since reliabilities for base statements are clearly dependent, if only because inferred from the same data, exact computation is infeasible. However, by utilizing ideas from the area of imprecise probability theory, one can compute bounds on the reliabilities of inferred statements. Specifically, in this chapter Fréchet inequalities are used to estimate these bounds. Moreover, two different variants of a bound estimate algorithm are considered: greedy and delayed. In simulation experiments, the delayed variant, at the expense of more book-keeping and computation time, does provide slightly tighter intervals. The work of this method is also validated on a real-world data set about ADHD which was recently published (Sokolova et al., 2016b).

In Chapter 4 the pathway between ADHD and one of its candidate genes DAT1, encoding the dopamine transporter is discussed. In an attempt to clarify its functional relevance to ADHD, the brain activity during the reward anticipation phase of the Monetary Incentive Delay (MID) task in a functional MRI paradigm was assessed. The MID task activates the ventral striatum (a region key to reward processing), where DAT1 is most highly expressed. A previous analysis based on standard statistical techniques did not show any significant
dependencies between a variant in the DAT1 gene and brain activation (Hoogman et al., 2013). In this study causal modeling was used for data analysis. The Bayesian Constraint-based Causal Discovery (BCCD) algorithm (Claassen and Heskes, 2012a) is able to find direct and indirect dependencies between variables, determines the strength of the dependencies, and provides a graphical visualization to interpret the results. Through BCCD one gets an opportunity to consider several variables together and to infer causal relations between them. Application of the BCCD algorithm confirmed that there is no evidence of a direct link between DAT1 genetic variability and brain activation, but suggested an indirect link mediated through inattention symptoms and diagnostic status of ADHD. This finding of an indirect link of DAT1 with striatal activity during reward anticipation might explain existing discrepancies in the current literature. Further experiments should confirm this hypothesis. This work has also been published (Sokolova et al., 2015b).

In Chapter 5 the basis for the strong correlation between two symptom domains of ADHD, inattention and hyperactivity/impulsivity are investigated. Both dimensions show high internal consistency and moderate to strong correlations with each other. It is not clear, however, what drives this strong correlation. The aim of this chapter was to address this issue. Causal discovery was applied on three independent data sets of scores of the two ADHD dimensions in NeuroIMAGE, and IMpACT, assessed by different raters and instruments, and further used information on gender or a genetic risk haplotype. In all data sets strong statistical evidence was found for the same pattern: the clear dependence between hyperactivity/impulsivity symptom level and an established genetic factor (either gender or risk haplotype) vanishes when one conditions upon inattention symptom level. Under reasonable assumptions, e.g., that phenotypes do not cause genotypes, a causal model that is consistent with this pattern contains a causal path from inattention to hyperactivity/impulsivity. The robust dependency cancellation observed in three different data sets suggests that inattention is a driving factor for hyperactivity/impulsivity. This causal hypothesis can be further validated in intervention studies. The model suggests that interventions that affect inattention will also have an effect on the level of hyperactivity/impulsivity. On the other hand, interventions that affect hyperactivity/impulsivity would not change the level of inattention. This causal model may explain earlier findings on heritable factors causing ADHD reported in the study of twins with learning difficulties such that inattention could be viewed as the key moderator. This work has been published (Sokolova et al., 2016a).

In Chapter 6 we study the comorbidity of ASD and ADHD using causal discovery. A large phenotypic data set was used, to infer a structural equation model using a causal discovery algorithm. Three distinct pathways between
1.6. Outline of the thesis

ASD and ADHD were identified: (1) from impulsivity to difficulties with understanding social information, (2) from hyperactivity to stereotypic, repetitive behavior, (3) a pairwise pathway between inattention, difficulties with understanding social information, and verbal IQ. These findings may inform future studies on understanding the pathophysiological mechanisms behind the overlap between ASD and ADHD. This work has also been published (Sokolova et al., 2017).

Overall the thesis used the following data sets: NeuroIMAGE (Chapter 2,5) Reversal task study (Chapter 2), Impact study (Chapter 3, 4, 5), ADHD 200 (Chapter 5), IMAGE (Chapter 6), and BOA (Chapter 6).
Handling hybrid and missing data in constraint-based causal discovery to study the etiology of ADHD

Causal discovery is an increasingly important method for data analysis in the field of medical research. In this chapter we consider two challenges in causal discovery that occur very often when working with medical data: a mixture of discrete and continuous variables and a substantial amount of missing values. In this chapter we develop a new method that can handle these challenges based on the assumption that data is missing at random and that continuous variables obey a non-paranormal distribution. We demonstrate the validity of our approach for causal discovery on simulated data as well as on two real-world data sets from a monetary incentive delay task and a reversal learning task. Our results help in the understanding of the etiology of attention deficit-hyperactivity disorder (ADHD).

2.1 Introduction

In recent years, the use of causal discovery in the field of medical research has become increasingly popular. Causal discovery analyses all variables together and suggests causal dependencies between variables, providing better insight.
Chapter 2. Handling hybrid and missing data in constraint-based causal discovery to study the etiology of ADHD

into the data. This approach has several advantages in comparison to standard statistical techniques. First, causal discovery provides an opportunity to learn causes and effects from the observed data, without performing experiments that can be costly and time consuming. Second, it detects whether the dependency between variables is direct or mediated through other variables. Third, it can visualize the results in the form of a graph that makes the results easier to interpret.

Even though there are a variety of algorithms that can learn the structure of the causal network for medical data, there are still many challenges in this field of research. In this chapter we discuss two of them. The first challenge is dealing with data that contains a mixture of discrete and continuous variables. Medical data often contains both discrete and continuous variables, where continuous variables are not necessarily normally distributed. The second challenge is dealing with incomplete data. In practice some tests are performed only for part of the patients, the quality of some data is poor, participants drop out, etc.

Although there are methods that can handle mixed variables or missing values separately, to the best of our knowledge there is no algorithm that can handle both challenges simultaneously for directed graphical models. However, there are such methods for undirected graphical models. In (Liu et al., 2012; Abegaz and Wit, 2014; Wang et al., 2009a) the authors propose different methods to estimate the correlation matrix for data with missing values and mixture variables, and based on this correlation matrix learn the structure of the undirected graphical model.

Algorithms that search for a structure of directed and undirected graphical models have a lot in common. They both try to find the optimal structure that provides the lowest complexity and the best goodness of fit. The main difference is that one model gives as output a directed graph and another gives an undirected graph. In this chapter we propose to transfer the ideas of structure learning for undirected graphical models to causal discovery.

We propose a method that can handle missing values and mixture variables based on the ideas for undirected graphical models presented in (Wang et al., 2009a; Abegaz and Wit, 2014). This method relies on two main assumptions. The first assumption is that the part of the data with continuous variables obeys a so-called non-paranormal distribution. For univariate monotone functions $f_1, \ldots, f_d$ and a positive-definite correlation matrix $\Sigma^0 \in \mathbb{R}^{d \times d}$ we say that a $d$-dimensional random variable $X = (X_1, \ldots, X_d)^T$ has a non-paranormal distribution $X \sim \text{NPN}_d(f, \Sigma^0)$, if $f(X) = (f_1(X_1), \ldots, f_d(X_d)) \sim N_d(0, \Sigma^0)$. We further assume functions $f_1, \ldots, f_d$ to be strictly monotone enabling computational tractability of the nonparanormal. A non-paranormal distribution implies that observed variables have monotonic relationships. This comes from the fact that a
Gaussian distribution implies linear, hence monotonic relationships between the surrogate variables \( f_i(x_i) \). Moreover, the monotonic relationship from surrogate Gaussian variables \( f_i(x_i) \) to observed variables \( x_i \) does not change their ratings. That implies monotonic relationships between observed variables as well. For most real-world medical data this is a reasonable assumption, since medical data usually has a relatively small sample size and non-monotonic dependencies, if present, are difficult to detect. The second assumption is that data is missing at random (MAR). This is also a reasonable assumption for many medical studies where the missing data often occur due to the fact that some experiments finish faster than others. As a result, information about symptoms, age, gender is usually present for all patients at the beginning of the study, while information about genes or brain functioning takes years to be collected and then may be missing for some subjects.

We propose a three-step algorithm to tackle the problems of missing data and non-paranormal distribution on each step: 1. Transform initial data into a Gaussian distribution by transforming the data first to the empirical distribution and then to Gaussian normal scores. This step deals with a mixture of discrete and continuous variables with non-paranormal distribution. 2. Use the expectation maximization (EM) algorithm to estimate the correlation matrix for this data. This step deals with missing values. 3. Apply a causal discovery algorithm to learn the causal structure from the correlation matrix. In this chapter we use the Bayesian Constraint-based Causal Discovery (BCCD) algorithm (Claassen and Heskes, 2012a) which is a state-of-the-art algorithm for causal discovery. This step outputs the causal graph and provides a reliability measure for each edge in the graph.

In the first part of the algorithm we use a copula transformation to estimate the correlation between variables. It is used to solve the problem with non-Gaussian data. This approach has been shown to work well for variables with non-paranormal distributions (Wang et al., 2009a; Harris and Drton, 2013). In our case we apply the same approach for a mixture of discrete and continuous variables and model the distribution of both discrete and continuous variables using a Gaussian copula to obtain an approximation of the correlation matrix. This leads to a slight underestimation of some correlations (Hoff, 2007). In case the focus of the research is the causal directed acyclic graph (DAG) from the observed variables, conditional independencies involving discrete variables do not exactly correspond to conditional independencies between their surrogate Gaussian variables. In this chapter, we focus on independencies in the surrogate variables, and assume that our data comes from a causal DAG in the latent space. Following Abegaz and Wit (Abegaz and Wit, 2014) i.e., it might not be necessary to use complex methods to model discrete variables, since this
would not result in a significant increase in accuracy. Further in the simulation study, we demonstrate that using this approximation our algorithm manages to accurately estimate the causal structure.

We compare the first two steps of the proposed algorithm with alternative methods. For the first step instead of transforming data to a Gaussian we transform it to ranks. For the second step instead of EM we use pairwise correlation, list-wise deletion, and mean imputation. Although these methods rely on a stronger assumption than EM that data is missing completely at random (MCAR), we choose them as a common alternative to EM. We show that EM with Gaussian transformation performs better than the alternative methods, when the amount of missing data is significant. We also show that the strength of the dependencies in the data influence the method that should be used to estimate the correlation matrix for causal discovery. Thus, even though the methods that are considered in this chapter to estimate correlation matrices have similar performance for the undirected graphical model, our analysis suggests that these methods have a different effect on the accuracy of a causal discovery algorithm. To test the validity of our conclusions that EM with a Gaussian transformation performs better than alternatives for directed graphical models, we repeat the same experiments with the PC algorithm instead of BCCD.

As a prototypical example we apply the proposed algorithm to two data sets about attention deficit hyperactivity disorder (ADHD). ADHD is a frequent and highly heritable neuropsychiatric disorder, affecting 5-6% of children (Polanczyk et al., 2007). Symptoms persist into adulthood in up to 50% of the childhood cases (Faraone et al., 2006). ADHD is characterized by two types of symptoms: hyperactivity/impulsivity and inattention, which can occur separately or combined. Given the large number of patients and long term impact of the disorder on patients and health care system, ADHD is a serious financial burden to society.

Both ADHD data sets used in this chapter have all features of a typical medical data set since they describe causal relationships between various possible factors of the disease such as genes, age, gender, and different types of symptoms and behavioral characteristics. These data sets have several possible factors, which can influence symptoms and interact with each other. The first data set describing a monetary incentive delay task has a moderate sample size of 409 subjects and approximately 10% of missing data. The second data set describing a reversal task has a sample size of 271 subjects and 0.3% of missing data. Both data sets have a mixture of discrete and continuous variables.

These data sets are part of the NeuroIMAGE project (see www.neuroimage.nl), whose goal is to learn cognitive, neural (MRI, MRS), and genetic underpinnings of ADHD. The first data set (von Rhein et al., 2014)
investigates the role of the genetic factors on the ADHD symptoms, and brain functioning measured during the reward related task. The second data set studies how problems with learning from reinforcement are associated with ADHD symptoms using a probabilistic reversal learning task (PRL). Based on this data, we build two causal models that provide deeper understanding of the altered reward processing and reversal learning in adolescents with ADHD than standard statistical tests. These models can help to understand the mechanisms that drive ADHD and make treatment more effective.

The rest of the chapter is organized as follows. Section 2.2 describes background information about causal discovery and graphical models. Section 2.3 describes algorithms for structure learning. Section 2.4 explains the proposed method. Section 2.5 presents the results of the experiments on simulated data and ADHD data. Section 2.6 provides our conclusion and future work.

### 2.2 Background

A Bayesian network is a pair $(\mathcal{G}, \Theta)$ where $\mathcal{G} = (\mathcal{X}, \mathcal{E})$ is a DAG with a set of nodes $\mathcal{X}$ representing domain variables and a set of arcs $\mathcal{E}$; $\theta_{\mathcal{X}_i} \subset \Theta$ is a set of parameters representing the conditional probability of variable $\mathcal{X}_i \subset \mathcal{X}$ given its parents $\text{Pa}_i$ in a graph $\mathcal{G}$. Using Bayesian networks we can model causal relationships between variables. In that case an edge $A \rightarrow B$ between variables represents a direct causal link from $A$ to $B$. This means that $A$ influences the values of $B$, but not the other way around.

Saying that two variables $A$ and $B$ are conditionally independent given $C$, means that if we know $C$, learning $B$ would not change our belief in $A$. Two DAGs are called equivalent to one another, if they entail the same conditional (in)dependencies. All DAGs that are equivalent to a graph $\mathcal{G}$ form an equivalence class of a graph $\mathcal{G}$, where all members are indistinguishable in terms of implied independencies. To represent the members of this equivalence class, a different type of structure is used, known as a partially directed acyclic graph (PDAG).

The three main assumptions that are often used when learning the structure of causal networks are the following (Spirtes et al., 2000):

1. **Causal Markov Condition**: each variable is independent of its non-descendant conditioned on all its direct causes.

2. **Faithfulness assumption**: there are no independencies between variables that are not implied by the Causal Markov Condition.
Chapter 2. Handling hybrid and missing data in constraint-based causal discovery to study the etiology of ADHD

3. Causal sufficiency assumption: there are no common confounders of the observed variables in $\mathcal{G}$ that are not members of the set.

In this chapter, we do not rely on the causal sufficiency assumption, i.e., we do allow for latent variables. One can represent the structure of a Bayesian network with latent variables using a so-called Maximal Ancestral Graph (MAG) (Richardson and Spirtes, 2002) on only the observed variables. In contrast to DAGs, MAGs can also contain bi-directed $X \leftrightarrow Y$ arcs (indicating that there is a common confounder) and undirected arcs $X \rightarrow Y$ (meaning that there is a selection bias affecting $X$ and $Y$). The equivalence class for MAGs can be represented by a partial ancestral graph (PAG) (Zhang, 2008). Edge directions are marked with “$-$” and “$>$” if the direction is the same for all MAGs corresponding to the PAG and with “$\circ$” otherwise.

2.3 Related study and motivation

In this section, we discuss existing methods for causal discovery. Since there are no algorithms that can handle mixture variables and missing data simultaneously, we consider the methods that can handle at least one of the challenges. Then we discuss how both challenges are solved for undirected graphical models and in Section 2.4 propose how can we transfer these ideas to directed models.

2.3.1 Structure learning

Causal discovery requires structure learning for directed graphical models. There are many methods that can be used to learn the structure of directed graphical models. A broad description of methods can be found in (Daly et al., 2011). In general, methods are divided into two approaches: constraint-based and score-based. The constraint-based approach works with statistical independence tests. First, this approach finds a skeleton of a graph by starting from the complete graph and excludes edges between variables that are conditionally independent, given some other set of variables (possibly empty). Second, the edges are oriented to arrive at an output graph. The constraint-based approach learns the equivalence class of DAGs and outputs a PDAG. Examples of the constraint-based approach are the IC algorithm (Pearl and Verma, 1991), PC-FCI (Spirtes et al., 2000), and TC (Pellet and Elisseeff, 2008). The score-based approach uses a scoring metric. It measures the data goodness of fit given a particular graph structure and accounts for the complexity of the network. There are many different scoring metrics, where the Bayesian score (Dawid, 1984) and the BIC score (Schwarz, 1978) are among the most common. The goal is to find
the graph that has the highest score. Unfortunately, this optimization problem is NP-hard, so different heuristics are used in practice. These methods are divided in local search methods, such as greedy search (Chickering et al., 1995), greedy equivalence search (Chickering, 2002), and global search methods, such as simulated annealing de Campos and Huete (2000) and genetic algorithms (Larrañaga et al., 1996). An algorithm by (Lemeire et al., 2012) is an example of an algorithm that does not rely on the faithfulness assumption.

An advantage of the constraint-based approach is that it does not have to rely on the causal sufficiency assumption, which means that the algorithm can detect common causes of the observed variables. A disadvantage of the constraint-based approach is that it is sensitive to propagating mistakes in the resulting graph. A standard approach makes use of independence tests, which results for borderline independencies/dependencies sometimes can be incorrect. The outcome of learning a network can be sensitive to such errors. In particular, one such error can produce multiple errors in the resulting graph. A set of conservative methods such as conservative PC (CPC) (Ramsey et al., 2006) and conservative FCI (CFCI) (Spirtes et al., 2004) tackles the problem of lack of robustness, outperforming standard constraint-based methods such as PC. An advantage of the score-based approach is that it provides a measure of reliability of inferred causal relations. This makes the interpretation of the results easier and prevents incorrect categorical decisions. A main drawback of the approach is that it relies on the causal sufficiency assumption and as a result cannot detect latent confounders.

To deal with a mixture of discrete and continuous variables several methods have been proposed for constraint-based structure learning. Spirtes et al. (2000) proposed to use conditional independence tests based on partial correlation. Harris and Drton (2013) showed that substituting Pearson correlation by Spearman correlation, the PC algorithm is able to infer a correct network structure under the assumption that data obey a Gaussian copula distribution. Margaritis (2004) developed a conditional independence test that does not rely on the distribution of the variables, but the test still involves discretization of the variables. Several methods have been proposed for score-based methods that can work with a mixture of discrete and continuous variables. Geiger and Heckerman (1994) proposed a closed-form solution for the Bayesian score of a mixture of discrete and continuous variables, but this solution only works in case a number of assumptions are met. These assumptions imply that the data are drawn from a conditional Gaussian distribution and forbid structures in the network with a continuous variable having a discrete variable as a child.

An alternative method is described in (de Santana et al., 2007) which uses a multiple regression framework for scoring structures. However, the method is
applicable only for time-series data. Bach and Jordan (2002) use Mercer kernels to estimate the structure of causal models, but calculation of a Gramm matrix requires significant computational costs ($O(N^3)$, where $N$ is the sample size) and may be inefficient for data sets with large sample sizes. Monti and Cooper (1997) use neural networks to represent the density function for a mixture of discrete and continuous variables. Estimation of the neural network parameters requires significant computational costs which makes this approach computationally expensive.

To deal with missing values several methods have been proposed to learn the structure of the network in the presence of missing values. Friedman (1998) proposed a Structural EM algorithm to estimate a Bayesian Network that has been further developed by Bernardo et al. (2003). The disadvantage of the EM algorithm is that it can get stuck in a local minimum. To prevent this, an evolutionary algorithm in combination with MCMC was proposed in (Riggelsen and Feelders, 2005). The limitation of these algorithms is that they usually rely on the assumption that data is either discrete or continuous Gaussian. Thus, current methods solve one of the problems at a time, either deal with missing value or with mixture variables. In this Chapter we propose a method that can handle both problems.

### 2.3.2 Undirected graphical models

In this section we present algorithm to find a structure for undirected graphical models for the data with mixture variables and missing data that inspired the solution for directed graphical models. Undirected graphical models build a graph where nodes represent variables and edges describe conditional independence relationships between the variables. The conditional independence relationships are estimated using the precision matrix (the inverse of a covariance matrix). Assuming that the precision matrix is sparse, the sparseness constraint is incorporated in the estimation of the precision matrix. That results in an optimization problem (Banerjee et al., 2008) to find the inverse correlation matrix $\Theta = \Sigma^{-1}$ with the best combination of goodness of fit and sparsity:

$$\max f(\Theta) = \log \det \Theta - \text{tr}(S\Theta) - \lambda \|\Theta\|_1.$$  \hspace{2cm} (2.1)

Here $\text{tr}$ denotes the matrix trace, $\det$ denotes determinant, $\|\Theta\|_1$ denotes the $L_1$ norm, $S$ denotes the empirical covariance matrix, and $\lambda > 0$ is a regularization parameter. In some sense, score-based structure learning algorithms for directed graphical models solve a similar optimization problem, but produce a directed graph as output.
In recent years, considerable effort has been invested in estimating the structure of undirected graphical models for non-Gaussian data and data containing missing values (Abegaz and Wit, 2014; Wang et al., 2009a). The precision matrix can be estimated under the assumption that data obey a non-paranormal distribution. In that case, Pearson correlation, which relies on the assumption of Gaussian data, is substituted by Spearman (Rho) rank correlation ($\rho$) (Liu et al., 2012; Abegaz and Wit, 2014; Wang et al., 2009a). An adjustment to the final Spearman’s rho correlation is applied in order to make it close to the Pearson correlation matrix, when the data is indeed Gaussian (Kendall, 1948; Kruskal, 1958):

$$S = 2 \sin(\pi \rho / 6).$$

The precision matrix can still be estimated when there are missing values in the data. One can use pairwise analysis and calculate pairwise correlation instead of complete case correlation to estimate the matrix (Abegaz and Wit, 2014; Wang et al., 2009a). As a result, one can keep as much data as possible. Another advantage of the pairwise correlation is that it does not introduce any bias to the results in contrast to imputation methods. However, there is no guarantee that the correlation matrix will be positive definite when we use pairwise correlation for data with missing values. In that case, a projection to the closest positive definite correlation matrix can be made (Boyd and Xiao, 2005; Higham, 1988).

Alternatively, the expectation maximization (EM) algorithm can be used to estimate the values of the correlation matrix $\Sigma$ (Little and Rubin, 1987; Dempster et al., 1977). The EM algorithm requires Gaussian data, so a copula transformation to Gaussian data can be used. The EM algorithm guarantees that the matrix would be positive definite, so no further adjustments are required.

Using Spearman pairwise correlation or the EM algorithm in combination with an optimization subroutine like Glasso or DoPing, showed to be one of the best methods in the field of undirected graphical models to estimate the structure of the graph with data obeying a non-paranormal distribution and missing values (Wang et al., 2009a). In this chapter we transfer these ideas to learn the structure of a causal graph and compare different methods using simulated and real-world data.

### 2.4 Proposed method

In this section we propose a causal discovery algorithm that can deal with both a mixture of discrete and continuous variables as well as missing data. In the first two steps of this algorithm we estimate the correlation matrix, when the data has mixture variables and missing data, based on the ideas described in Section 3. In
the third step, we use this correlation matrix as an input into a causal discovery algorithm to infer the causal structure. We use the BCCD algorithm for this purpose, one of the state-of-the-art algorithms in causal discovery. Claassen and Heskes (2012a) showed that BCCD outperforms reference algorithms in the field, such as FCI and Conservative PC. Moreover, it provides an indication of the reliability of the causal links that makes it easier to interpret the results and compare alternative models. The advantage of the BCCD algorithm is that it combines the strength of constraint-based and score-based approaches. We rely on the assumption that data is missing at random and that continuous variables obey a non-paranormal distribution. This is a valid assumption for many medical data sets as has been discussed in Section 2.1.

We propose the following algorithm:

**Step 1 Mixture of discrete and continuous variables**

To deal with data sets that contain a mixture of discrete and continuous variables we propose to use a Gaussian copula. For each variable \( X_i \) in the data set we estimate the rescaled empirical distribution

\[
\hat{F}_i(x) = \frac{1}{n+1} \sum_{j=1}^{n} \mathbb{I}\{X_{i,j} < x\},
\]

where \( \mathbb{I} \) is an indicator function and then transform the data into Gaussian normal scores

\[
\hat{X}_i = \Phi^{-1}_i(\hat{F}(X_i)).
\]

In this step missing values are ignored.

**Step 2 Correlation matrix with missing data**

The next step is to estimate the correlation between the variables in the model. This correlation matrix will be used in the next steps, where we will estimate the causal structure of the graph. New variables now have a Gaussian distribution, so we can use Pearson correlation to estimate dependencies between variables. Since our data has missing values, we propose to first use the EM algorithm to estimate the correlation matrix, since this algorithm provides an unbiased estimate of parameters and their standard error (Dempster et al., 1977).

The EM algorithm searches for the Maximum Likelihood Estimate (MLE) of the marginal likelihood by iteratively applying the following two steps:

1. E-step: Estimate the sufficient statistics;
2. M-step: Re-estimate the covariance matrix using the sufficient statistics from the previous step. Re-estimate missing values. The algorithm iterates until convergence.

The output of EM is a covariance matrix that should be normalized to have unit variance.

**Step 3 Apply BCCD**

The correlation matrix is used in the BCCD algorithm to estimate the causal structure of the graph. We here describe only the basic idea of the BCCD algorithm. A more detailed description can be found in (Claassen and Heskes, 2012a). The BCCD algorithm contains two main steps:

**Step 3.1** Start with a fully connected graph and perform adjacency search, estimating the reliability of causal relations, for example $X \rightarrow Y$. If a causal relation declares that variables are conditionally independent with a reliability higher than a predefined threshold, delete an edge from the graph between these variables. To estimate the reliability of the causal statement, we have to do the following substeps repeatedly:

(a) First we estimate the mutual information, using the correlation matrix $\Sigma$ that we get as an output from Step 2. We propose to use the following formula:

$$I(X_i, X_{Pa_i}) = -\frac{1}{2} \log \frac{|\Sigma_{i,Pa_i}|}{|\Sigma_{Pa_i}|}, \quad (2.5)$$

where $X_{Pa_i}$ are the parents of node $i$ in DAG $\mathcal{G}$, $\Sigma_{Pa_i}$ is a correlation matrix between the parents of variable $X_i$, and $\Sigma_{i,Pa_i}$ is a correlation matrix between variable $X_i$ and its parents.

(b) Knowing the value of mutual information we can estimate the Bayesian Information Criterion (BIC) for data $\mathbf{D}$ that can then be used to compare scores of different DAGs ($\mathcal{G}$). The BIC score is decomposed into the sum of two components, the mutual information $I(X_i, X_{Pa_i})$ estimated in the previous substep and $\text{Dim}[^{\mathcal{G}}]$ the number of parameters necessary to estimate the model.

$$\text{BICscore}(\mathbf{D}|\mathcal{G}) = M \sum_{i=1}^{n} I(X_i, X_{Pa_i}) - \frac{\log M}{2} \text{Dim}[^{\mathcal{G}}], \quad (2.6)$$

27
where \( n \) is the number of variables, and \( M \) is the sample size. The first component measures the goodness of fit, and the second penalizes the complexity of the model.

(c) To estimate the reliability measure, we need to estimate the marginal likelihood \( p(D|G) \). We propose to use BIC, which approximates the logarithm of the marginal likelihood:

\[
\log p(D|G) = \text{BICscore} + O(1). \tag{2.7}
\]

To get the probability \( p(D|G) \), we should calculate (2.7) for all possible graphs for this subset of variables and then normalize it.

(d) Now we can estimate the reliability of the causal statement \( L \), e.g., \( L : 'X \rightarrow Y' \). It gives a conservative estimate of the probability of a causal relation. We estimate the reliability measure using a Bayesian score:

\[
p(L|D) = \frac{\sum_{M \in M(L)} p(D|M)p(M)}{\sum_{M \in M} p(D|M)p(M)}, \tag{2.8}
\]

where \( p(D|M) \) denotes the probability of data \( D \) given structure \( M \), \( p(M) \) represents the prior distribution over structures, and \( M(L) \) is the set of structures containing the relation \( L \). In this equation we approximate the probability \( p(D|M) \) by \( p(D|G) \), which was calculated in the previous substep. Equation (2.8) also requires to set the prior distribution for \( p(M) \). Claassen and Heskes (2012a) propose to use a uniform prior. More details can be found in (Claassen and Heskes, 2012a).

**Step 3.2** Rank all causal relations in decreasing order of reliability and orient edges in the graph starting from the most reliable relations. If there is a conflict, pick the causal relation that has a higher reliability.

To estimate Equation (2.8) in Step 3.1, the algorithm requires calculating the marginal likelihood over all possible graphs for each causal relation that we infer. For speed and efficiency of the algorithm, the set of possible graphs is limited to the graphs with at most five vertices, which gives a list of at most 29,281 DAGs per set of five variables (Claassen and Heskes, 2012a) to reduce the computational complexity. In theory, limiting the number of vertices to five may lead to a loss of information. In practice, however, the accuracy of the BCCD algorithm is hardly affected and it still outperforms standard algorithms that perform conditional independence tests for more than five variables (Claassen and Heskes, 2012a).
In our method we assume that each observed variable has a corresponding latent, surrogate variable, with a monotonic relationship between the two. The latent variable can thus be seen as a surrogate value, representing the exact same concept as the corresponding observed variable. The method infers, and then depicts in the output graph the causal structure between these surrogate variables. This relates to our assumption of nonparanormal distribution.

Each step in the proposed algorithm has several possible alternative solutions. Based on the papers about undirected graphical models (Wang et al., 2009a; Abegaz and Wit, 2014), an alternative for Step 1 is to transform data to ranks and use Spearman to deal with mixture variables. To deal with missing variables in Step 2 we use either pairwise correlation, mean imputation, or listwise deletion to deal with missing values. In case of pairwise correlation there are no guarantees that the correlation matrix will be positive definite and if not it should be projected to the closest positive definite matrix. Calculating Spearman pairwise correlation we have two alternatives: to apply the transformation proposed in Equation (2.2) or not to apply.

An alternative to Step 3 could be any score based causal discovery algorithm that can use a correlation matrix as an input. In this chapter we focus on the alternatives for Steps 1 and 2 and would like to learn which approach is the best for directed graphical models. Thus, we do not try to find the best alternative for Step 3, but rather check whether the best approach for Step 1 and 2 is the same when we use a different causal discovery algorithm. In order to do so, we compare our results with the PC algorithm.

2.5 Experimental results

2.5.1 Simulation study

To estimate the accuracy of the causal discovery for different alternatives of Steps 1 and 2 of the algorithm discussed in the previous section, we made a simulation study. We chose the Waste Incinerator Network (Lauritzen and Lauritzen, 1992) which contains a mixture of discrete and continuous variables. The Waste Incinerator Network describes the emission from a waste incinerator depending on the filter efficiency, waste type, burning regime, and other factors. The network contains nine variables that are connected by ten arcs as can be seen in Figure 2.1.

The original version of the network contains continuous Gaussian variables. To make these variables nonnormal (here we mean non-Normally distributed), we applied a monotonic transformation ($X^3$). We considered the Waste Incinerator Network when the correlation between variables is extreme-high (the cor-
relation matrix is close to singular) and medium (the parameters that were used are provided in supplementary material). We generated data with three levels of missing data (0%, 5%, and 30%) and four sample sizes: 100, 250, 500, and 1000. We repeated our experiments 50 times. Performance was measured by the PAG accuracy measure that evaluates how many edges were oriented correctly in the output PAG compared with the ground-truth PAG (Figure 2.1(b)). We also estimated the correctness of the skeleton by calculating precision and recall metrics, where the former estimates the number of edges inferred correctly to the total number of inferred edges and the latter estimates the number of edges inferred correctly to the number of edges in the ground truth graph (Figure 2.1).

We investigate the effect of different approaches to estimate the correlation matrix (described in Steps 1 and 2 in Section 4) on the accuracy of the causal discovery algorithm. We consider the following alternatives:

1. Pearson correlation with EM. (EM)

2. Spearman correlation with mean imputation. (Spearman mean)

3. Spearman correlation with list-wise deletion. (Spearman list-wise)

4. Pairwise Spearman correlation. In this approach we do not make an adjustment of the Spearman correlation based on (2.4). (Spearman not adjusted)

5. Pairwise Spearman correlation with adjustment. In this approach we do make an adjustment of the Spearman correlation based on (2.4). (Spearman pairwise)
If the obtained matrix is not positive definite, it is projected to the closest positive definite matrix (Higham, 2002). We repeat these tests for two different causal discovery algorithms: BCCD and PC.

When there is no missing data, Spearman mean, Spearman list-wise, and Spearman pairwise provide the same results. Thus, we compare only three alternatives: EM, Spearman, Spearman adjusted. Figure 2.2 represents the results of BCCD for two cases: when the data has a medium correlation and high correlation. For medium correlation, Spearman adjusted performs similarly with the other two methods, but for high correlation it performs significantly worse than Spearman not adjusted and EM. The factor that is causing this difference is the ill-defined determinant of the correlation matrix which is close to zero when the correlation is high. Adjustment of the correlation matrix using (2.4) increases the correlations even more, which results in a non-positive definite correlation matrix and loss of conditional independencies between variables compressed in the correlation matrix. This results in many incorrect edges and a low PAG accuracy. Thus, when the correlation between variables is high, adjusting the Spearman correlation may lead to significantly worse results. Based on this conclusion we did not consider Spearman adjusted for tests with missing values, since it already showed significantly worse performance compared to Spearman not adjusted.

Figure 2.3 shows the results of BCCD when the data have a low (5%) and high (30%) percentage of missing values. When percentage of missing values is low (5%) the differences between EM, Spearman mean, Spearman list-wise, and Spearman pairwise are not significant. When the percentage of missing values is high (30%), EM performs significantly better than Spearman for both medium and high correlation. One of the main factors that leads to this difference in performance between EM and pairwise correlation is a non positive definite correlation matrix with a high number of missing values. The advantage of the EM algorithm in that case is that it outputs a positive definite matrix. Even though we projected the Spearman correlation matrix to a positive definite correlation matrix, simulation tests show that EM provides more accurate results. When the percentage of missing values is high, mean imputation leads to a decrease in variance which results in lower accuracy. As expected, Spearman list-wise performs worse than all other methods due to significant loss of information when estimating the correlation when the amount of missing data is high.

We repeated the same experiments with PC and obtained similar patterns, see Figure 2.4. When 5% of the data is missing, no significant difference between the methods is observed. When 30% of the data is missing, EM gives significantly better PAG accuracy. Although BCCD is a more advanced algo-
Figure 2.2: The accuracy of the BCCD algorithm (PAG accuracy, precision, and recall) for the Waste Incinerator Network for data with medium and high correlation when there are no missing values.

Algorithm than PC it provides lower PAG accuracy in these experiments. It happens because PC infers the directions based on the assumption that there are no unobserved common causes and no selection bias, while BCCD does not rely on these assumptions. Since waste incinerator network does not contain unobserved common causes and selection, PC can infer the correct structure of the network more easily than BCCD. For both BCCD and PC increasing the sample size improves recall and PAG accuracy, while it does not help to improve the precision. When sample size becomes large our method starts to detect more spurious edges leading to a decrease in precision in the simulation studies. An increase in the number of spurious edges with an increase in sample size is a common problem in structure learning, since with a high sample size even very small correlations between variables become significant. In this case we are not talking about ‘spurious’ correlations (which would be resolved with more data), but about real but weak correlations that are often present in complex, real-world systems, but that are overlooked (not detected) in small data sets.

We compare our results with the results obtained for undirected graphical models in (Abegaz and Wit, 2014; Wang et al., 2009a). The two main results for undirected graphical models are: 1) Spearman and EM both perform well,
Figure 2.3: The accuracy of the BCCD algorithm (PAG accuracy, precision, and recall) for the Waste Incinerator Network for data with medium and high correlation at two levels of missing values: 5% missing, 30% missing.
Figure 2.4: The accuracy of the PC algorithm (PAG accuracy, precision, and recall) for the Waste Incinerator Network for data with medium and high correlation at two levels of missing values: 5% missing, 30% missing.
2.5. Experimental results

while EM performs slightly better. 2) Making the projection for the correlation matrix to the closest positive definite matrix improves the results. The main results that we obtained for directed graphical models are: 1) EM performs significantly better than Spearman with projection for data with a high percentage of missing values and a high correlation between variables. 2) Working with directed graphical models one should be careful in applying the adjustment of the Spearman correlation. This adjustment may destroy the positive definiteness property of the matrix even when there are no missing values in the data. The difference in results between undirected and directed graphical models can arise because undirected graphical models are typically inferred under sparseness constraints. Optimizing the correlation matrix under sparseness constraints decreases the number of spurious dependencies that might otherwise arise due to an ill conditioned or even non-positive definite correlation matrix. We do not have a similar type of regularization to estimate the mutual information in (2.5) and (2.6), which then may explain the larger difference in performance between EM, Spearman, and Spearman adjusted.

2.5.2 ADHD data

We have applied the BCCD algorithm with EM to two data sets representing two different ADHD studies performed as a part of the NeuroIMAGE study.

MID tasks study

The first study (von Rhein et al., 2014) investigated the brain response during reward anticipation and receipt with a monetary incentive delay (MID) task in a large sample of adolescents and young adults with ADHD, their unaffected siblings and healthy controls. All subjects participated in cognitive testing and neuroimaging. The brain activation was measured in ventral striatum (VS) and orbital-frontal cortex (OFC) brain areas during the reward anticipation and receipt (von Rhein et al., 2015). The data set contained 409 participants: 189 probands with ADHD, 104 unaffected siblings, and 116 age-matched controls. Since the presence of the unaffected siblings can blur the effect of the genes, we did not include them in our study and consider only ADHD patients and healthy controls. Approximately 10% of data is missing for this study. The main reason for the presence of missing values in this data set was that part of the experiments was very time consuming and as a result not all the results were available yet, leading to missing values in the data set. Thus, we may assume for this data set that data is missing completely at random. Scatter plots did not reveal any non-monotonic dependencies, supporting our hypothesis of monotonic
dependencies.

Using BCCD we wanted to infer the endophenotypic model (Franke et al., 2009) that explains the relationships between genes, brain functioning, behaviors, and disease symptoms. To apply causal discovery to this data set, domain experts selected 12 variables. These variables include general characteristics, genetic factors, comorbid disorders, symptoms, and results of the MID task experiments:

1. Gender (male/female)
2. Age
3. IQ
4. DAT1 risk gene (present/not present)
5. NOS1 risk gene (present/not present)
6. Inattention symptoms (score assessed by KSADS and CPRS-R:L)
7. Hyperactivity/impulsivity symptoms (score assessed by KSADS and CPRS-R:L)
8. Aggression (presence/absence of Oppositional Defiant Disorder (ODD) or Conduct disorder (CD))
9. Brain activation in OFC during receipt (Receipt OFC)
10. Brain activation in VS during receipt (Receipt VS)
11. Brain activation in OFC during anticipation (Anticipation OFC)
12. Reaction time difference (the difference in reaction time with and without a reward)

The initial data set contained two different estimates of the ADHD symptoms: one estimated by parents and another one estimated by a psychiatrist. Since these are highly correlated it makes no sense to include both. We decided to keep the parent scores, because an initial analysis revealed slightly more variation and slightly stronger correlation with the other variables. These symptom scores represent the quantiles in the population adjusted by age and gender. We readjusted these scores to be able to see the explicit effect of gender.

Partially due to the small sample size, the BCCD algorithm inferred only the skeleton of the network, but not the direction of the edges for the resulting
network. However, including prior knowledge about the domain that no variable in the network can cause gender, and the endophenotypic assumption from (Franke et al., 2009) that symptoms are the consequence of problems with brain functioning, BCCD inferred the direction of several edges.

The causal network learned from the data is presented in Figure 2.5. The figure indicates network edges with an estimated link of 50% or above. The resulting network structure provides an endophenotypic model that connects genes, brain functioning, and symptoms together. The causal model suggests association of genes with brain activation during the monetary incentive delay task. This model confirms several causal pathways that were previously presented in other studies, and suggests new endophenotypic pathways.

Our causal model suggests that NOS1 is associated with brain activation in OFC during reward receipt and DAT1 with brain activation during reward anticipation. The effect of genes on brain functioning was also claimed in other studies (Hoogman et al., 2011; Dreher et al., 2009). The model proposes that the reaction time depends on the age of the subject and his/her level of inattention. In (Hodgkins, 1963) a similar conclusion was drawn about the increase of reaction time up to early adulthood. The level of inattention symptoms depends on the gender of the subject. This statement is confirmed by different studies in the field of ADHD (Bauermeister et al., 2007). The level of hyperactivity/impulsivity depends on the level of inattention and on the problems with brain activation in MID task in VS. The effect of inattention on hyperactivity/impulsivity was also found in (Willcutt et al., 2000). The level of aggression is associated with the level of IQ and inattention level.

Most studies focus on association between symptoms and reward anticipation rather than between symptoms and reward receipt and several studies report a link between these two variables (Plichta and Scheres, 2014; Scheres et al., 2007), whereas others do not (Paloyelis et al., 2012). The causal model inferred in our study suggests a causal path from reward receipt to hyperactivity/impulsivity symptoms and no clear link between reward anticipation and symptoms. Moreover, the causal model provides computational evidence for new causal association between genes, brain functioning, and symptoms, from NOS1 to hyperactivity/impulsivity symptoms through brain functioning during receipt. The model inferred in this study should be treated with care, but can suggest further studies, zooming in on some of the pathways found through this analysis.

Reversal task study

The second study investigated the behavioral response during a probabilistic reversal learning task (PRL). With the PRL one can learn whether participants
are able to adapt to a changing situation, whether they are able to learn a (new) rule, and possibly whether participants are sensitive to reward and punishment. The participants of the reversal task study partially overlap with the participants from the MID task study. However, since the MID task experiments were performed several years before the reversal task study, in the reversal task study the participants are older.

We applied BCCD to investigate the relationships between ADHD symptoms and problems with reversal behavior. Based on the domain knowledge experts selected nine variables that are associated with ADHD and may influence the outcome of the reversal task:

1. Gender (male/female)
2. Age
3. IQ
4. Inattention symptoms (score assessed by KSADS and CPRS-R:L)
5. Hyperactivity/impulsivity symptoms (score assessed by KSADS and CPRS-R:L)
6. Win stay score (percentage of trials in which participants chose the same stimulus after a win)

7. Lose shift score (percentage of trials in which participants chose the other stimulus after a loss)

8. Preservative error score (the amount of errors made after reversal that were related to picking the previous stimulus)

9. Medication status (naive/not naive)

To infer a more accurate causal network we included in the model the prior knowledge that nothing can cause gender.

The causal network inferred by BCCD is presented in Figure 2.6. This network suggests the effect of age on subject’s IQ and whether the medication was prescribed or not. Moreover, age is associated with gender in this model, which happens due to age/gender unbalance in the sample. In contrast to the causal model in the MID task (Figure 2.5), this causal model does not find any link between gender and symptoms. A possible explanation can be the observation (Kooij et al., 2005) that gender unbalance vanishes when ADHD patients get older and become adults. Since in the reversal task study subjects are approximately 3.6 years older than in the MID task study, this might explain why in reversal task study there is no effect of gender on symptoms.

Analysis of the causal links between symptoms PRL experiment outcomes suggest that IQ and hyperactivity/impulsivity are associated with variables related to reversal learning. Subjects with a lower IQ and higher level of hyperactivity/impulsivity have a higher percentage of lose-shift responses and a lower percentage of win-stay responses, suggesting sensitivity for punishment but not for reward in participants with more hyperactivity/impulsivity symptoms. Although we did not find a direct association between symptoms and age, older participants with ADHD tend to have less hyperactivity/impulsivity symptoms than younger participants (Faraone et al., 2015), possibly relating age to performance in the PRL. Probably a sample with higher age differences is needed to be able to infer such a pattern from the data.

The association of IQ with both win-stay and lose-shift may be related to the difficulty of the task in general. Participants with a lower IQ have more problems with performing the task but are not specifically more sensitive to punishment than reward. Additionally, with the PRL one can investigate how well people can adapt to a changing rule, which may be difficult for subjects with ADHD (Abouzari et al., 2015). Although this is the first study of causal analysis with ADHD and PRL performance, it shows promising possibilities for future research.
Figure 2.6: The causal graph representing causal relationships between variables for the reversal task ADHD data set. The graph represents a PAG, where edge directions are marked with “−” and “− >” for invariant edge directions and with “− ◦” for non-invariant edge directions. The reliability of an edge between two variables is depicted with a percentage value near each edge.

2.6 Discussion and conclusions

The simulation study shows that the EM algorithm performs better than Spearman with pairwise correlation, mean imputation, and list-wise deletion for directed graphical models when the percentage of missing values is high, while providing similar results when the percentage is low. Comparing EM with pairwise Spearman these results can be explained by the fact that the correlation matrix can become non-positive definite when calculating pairwise correlation with missing values. This leads to an incorrect estimate of the determinant. Estimation of the correlation matrix using the EM algorithm outputs a positive definite matrix that results in a better accuracy of the algorithm. Thus the EM with Gaussian transformation proposed in this chapter performs better than the Spearman pairwise correlation method proposed in (Wang et al., 2009a) for causal discovery. EM outperforms mean imputation due to a more sophisticated method to impute missing values that does not reduce variance. Bad performance of list-wise deletion when the percentage of missing values is high is logical, since the main part of the data is not used when applying this method. A simulation study using the PC algorithm instead of BCCD confirmed these results. Although the EM algorithm is computationally more expensive than alternative methods described in the chapter such as pairwise correlation, it should be calculated only once. For a data set of 15 variables it does not take longer than a
minute.

Where pairwise and list-wise deletion correlation estimation that rely on the assumption that data is missing completely at random (MCAR), EM assumes that data is “just” missing at random (MAR). This assumption applies more often in practice and thus increases the range of data sets for which it can be used.

The simulation study also shows that adjustment of the Spearman correlation when the correlation is high can decrease the accuracy of the causal discovery algorithm. The determinant of the correlation matrix is close to zero when the correlation between variables is high. When applying the adjustment of the correlation matrix, the correlation increases even more which can again result in a non-positive defined matrix determinant. For medium correlation, Spearman adjusted and Spearman not adjusted show similar accuracy. Thus, we can conclude that for estimating mutual information it is better not to adjust the Spearman correlation.

Using the BCCD algorithm we inferred an endophenotypic model of ADHD during the MID task. The resulting model explains the effect of genes on brain functioning, the effect of brain functioning and general factors on disease symptoms, and an interaction between these symptoms. This model confirms previous findings in the literature and proposes new causal links between variables. The model shows evidence for receipt and against anticipation endophenotypes and highlights the need to extend genetic research on this less expected endophenotype. In this sense this study suggests promising new pathways for genetic research in ADHD that need to be confirmed by genetic imaging studies.

BCCD inferred a model explaining the interaction between symptoms and problems with reversal learning measured during the PRL task. This model suggests that the main factors that influenced the outcome of the experiments were hyperactivity/impulsivity, IQ, and medication. These results provide a new insight into the reversal learning problems and can improve its treatment.
Causal discovery provides an opportunity to infer causal relationships from purely observational data and to predict the effect of interventions. Constraint-based methods for causal discovery exploit conditional (in)dependencies to infer the direction of causal relationships as was explained in Chapter 2. They typically work through forward chaining: given some causal statements, others can be inferred by applying relatively straightforward causal logic such as transitivity and acyclicity. Starting from the premise that we can estimate reliabilities for base causal statements, in this chapter we propose a new approach to estimate the reliability of novel statements inferred by forward chaining. Since reliabilities for base statements are clearly dependent, if only because inferred from the same data, exact computation is infeasible. However, lending ideas from the area of imprecise probability theory, we can compute bounds on the reliabilities on inferred statements. Specifically, we make use of the good old Fréchet inequalities and discuss two different variants: greedy and delayed. In simulation experiments, we show that the delayed variant, at the expense of more bookkeeping and computation time, does provide slightly tighter intervals. We illustrate our method on a real-world data set about attention deficit/hyperactivity disorder.

3.1 Introduction

The use of causal discovery algorithms has become increasingly popular in recent years. Causal discovery algorithms are able to predict the result of an intervention under some reasonable assumptions, purely from observational data. Causal relationships between variables are typically represented through a causal directed acyclic graph (DAG), where a directed path from variable $X$ to $Y$ represents a causal path from $X$ to $Y$.

One of the most common approaches to learn a causal DAG from data is the so-called constraint-based approach. This approach employs a statistical independence test and typically consists of two steps. In the first step, the skeleton of the graph is learned based on conditional independencies inferred from the data. In the second step, edges of the graph are oriented. This orientation is based on the presence of so-called V-structures (or colliders). A V-structure is a triple $(X,Y,Z)$, where $Z$ is a common child of $X$ and $Y$, variables $X$ and $Y$ are independent, but they become dependent when we condition on variable $Z$. Based on the V-structures, causal statements are inferred, which determine the orientation of the edges. By combining causal statements, new statements can be inferred to orient other edges in the graph. Examples of popular constraint-based approaches are IC (Pearl and Verma, 1991) and PC/FCI (Spirtes et al., 2000).

If a causal statement is estimated incorrectly, this mistake may propagate through the whole graph, leading to many erroneous orientations. A potential remedy, suggested by, e.g., (Triantafilou et al., 2014; Claassen and Heskes, 2012b), is to keep track of the reliability of the inferred causal statements. Standard constraint-based methods consecutively apply hypothesis tests for conditional independence that provide a p-value as output. In (Triantafilou et al., 2014) authors propose to translate these p-values into probabilities, following the generic approach of Sellke et al. (2001), and use these probabilities to estimate the reliability of the skeleton. Claassen and Heskes (2012b) estimate the reliability of conditional independence statements by computing Bayesian scores on DAGs over subsets of variables, and combine these into reliabilities of both edges and orientations.

In constraint-based methods causal statements are typically inferred using forward chaining: consecutively combining earlier inferred statements. The causal statements themselves are clearly dependent, since inferred from the same data, which makes combination of their reliabilities highly nontrivial. In this chapter we propose to apply ideas from the theory of imprecise probabilities (Walley, 1991) or interval probabilities (Weichselberger, 2000) to keep track of reliability intervals instead of point estimates. By probability intervals we
3.2. A Method for Computing Causal Reliability Intervals

3.2.1 Constrained-Based Causal Discovery

In order to infer a skeleton and causal statements from data, we use the BCCD algorithm (Claassen and Heskes, 2012b). BCCD is a state-of-the-art algorithm for estimating causal relationships between variables that also provides a reliability measure for inferred statements and can handle both potential confounding (i.e., does not assume causal sufficiency) and selection bias. By selection bias we mean a process of data selection that introduces dependencies between variables that are not representative of the population. Confounding refers to an unobserved common cause between several variables. The output of BCCD is a so-called partial ancestral graph (PAG) (Richardson and Spirtes, 2003) that is used to represent DAGs with latent variables. In this chapter we provide a short description of BCCD, a more detailed description can be found in (Claassen and Heskes, 2012b) and in Chapter 2.

The BCCD algorithm consists of two main stages.

1. **Inference of the skeleton and base causal statements.** BCCD considers subsets of maximum K variables. For each subset, it computes a Bayesian score (Dawid, 1984) for every directed acyclic graph. Each directed acyclic graph implies particular causal statements of the form

\[
\begin{align*}
\text{no collider: } & (Z \Rightarrow X) \lor (Z \Rightarrow Y) \lor (Z \Rightarrow S) \\
\text{no causal path: } & (Z \nRightarrow X) \land (Z \nRightarrow S) \\
\text{causal path: } & (Z \Rightarrow X) \land (Z \nRightarrow S)
\end{align*}
\] (3.1)

assume an estimate of an interval in which future observations will fall. We develop a new method that estimates these probability intervals using the so-called Fréchet inequalities (Fréchet, 1935) and prove that a particular form of the logical statement gives the best lower and upper bound using Fréchet inequalities. Although in this chapter we focus on Bayesian Constraint-based Causal Discovery (BCCD) (Claassen and Heskes, 2012b), in principle our method can be applied to any constraint-based causal discovery algorithm that infers causal statements. We also propose an approximation of our algorithm that provides similar accuracy but requires lower computational complexity.

The rest of this chapter is organized as follows. In Section 3.2, we give the background and derive our methods for computing reliability intervals. In Section 3.3, we apply these methods to simulated and real-world data. In Section 3.4 we provide our conclusions and discuss future work.
where statement \((Z \Rightarrow S)\) indicates a selection bias \(S\) on variable \(Z\). The first line follows from a so-called minimal conditional independence: if conditioning upon a variable \(Z\) breaks the conditional dependence between two variables \(X\) and \(Y\), \(Z\) must have a causal path to \(X\), \(Y\), or both (more details can be found in (Claassen and Heskes, 2012b)). The second line follows from a minimal conditional dependence: if adding \(Z\) to the conditioning set makes a variable \(X\) dependent of another variable, say \(Y\), there cannot be a causal path from \(Z\) to \(X\) (nor from \(Z\) to \(Y\)) and there can be no selection bias \((S)\) on \(Z\). This corresponds to the V-structure mentioned before. The first line is in a sense the negation of the second line: it states that \(Z\) must be on a path between \(X\) and \(Y\), but cannot lead to a V-structure, so cannot be a collider on this path. BCCD infers the reliability for each causal statement by combining the Bayesian scores for all DAGs that match this statement. This reliability gives a conservative estimate of the probability of a causal relation (Claassen and Heskes, 2012b). We interpret it as an estimate for the probability that the causal statement is true.

2. Combination of causal statements. In the second stage, BCCD infers new causal statements by combining causal statements and applying rules from standard causal logic:

\[
\begin{align*}
\text{irreflexive:} & \quad (X \Rightarrow X) \vdash \text{false} \\
\text{acyclic:} & \quad (X \Rightarrow Y) \vdash (Y \nRightarrow X) \\
\text{transitive:} & \quad (X \Rightarrow Y) \land (Y \Rightarrow Z) \vdash (X \Rightarrow Z)
\end{align*}
\]

The system of causal statements is closed, in the sense that all newly inferred causal statements can also be written in the form (3.1), with special cases \((Z \Rightarrow X) \lor (Z \Rightarrow S)\) (causal path or selection bias; first line, when \(Y\) equals \(X\)) and \((Z \nRightarrow S)\) (no selection bias; second line, when \(X\) equals \(Z\)).

The output of the BCCD algorithm is a list of statements of the form \((X \Rightarrow Y),(Y \nRightarrow X),\) or \((X \Rightarrow S)\). Given a skeleton, these statements can be used to determine the directions of edges, e.g., a combination of two statements \((X \Rightarrow Y),(Y \nRightarrow X)\) suggests a causal effect of \(X\) on \(Y\) that is represented as \(X \rightarrow Y\) in a PAG. Statements \((X \Rightarrow Y)\), and \((Y \Rightarrow X)\) indicate a selection bias between \(X\) and \(Y\) and is represented with \(X \leftarrow Y\) in a PAG. If a list of statements contains \((X \nRightarrow Y),(Y \nRightarrow X)\) then there is a latent confounding between two variables that is represented as \(X \leftrightarrow Y\) in a PAG. Circle marks are used to mark edges which directions cannot be fully determined, e.g., if only statement \((X \Rightarrow Y)\) was inferred, it is either a causal effect \(X \rightarrow Y\) or a selection bias \(X \leftarrow Y\), which will be represented as \(X \leftarrow \circ Y\) in a PAG.

The parameter \(K\), which determines the subsets of variables considered by BCCD to infer causal statements, plays an important role in the BCCD algorithm. The higher \(K\), the more causal statements can be inferred directly, without the
need to combine statements using causal logic. On the other hand the complexity of the algorithm grows exponentially with \( K \). For example the number of possible causal models for which likelihoods should be estimated is 25 for \( K = 3 \) and 29.281 for \( K = 5 \). The default value of \( K \) is five variables, a fine compromise between complexity and accuracy. In this chapter, we will also consider \( K = 3 \), to better demonstrate the effect of different strategies for combining causal statements.

**Example 1.** As a running example, we will consider a so-called Y-structure (Mani et al., 2012). This structure, sketched in Figure 3.2.1, consists of four variables, where \( X_1, X_2, \) and \( X_3 \) form a V-structure and \( X_4 \) is a child of \( X_3 \). Given enough data generated from such a Y-structure, BCCD with \( K = 3 \) would infer the base causal statements.

\[
\begin{align*}
\psi_1 : (X_3 \Rightarrow X_1) \land (X_3 \Rightarrow S) \\
\psi_2 : (X_3 \Rightarrow X_2) \land (X_3 \Rightarrow S) \\
\psi_3 : (X_3 \Rightarrow X_1) \lor (X_3 \Rightarrow X_4) \lor (X_3 \Rightarrow S) \\
\psi_4 : (X_3 \Rightarrow X_2) \lor (X_3 \Rightarrow X_4) \lor (X_3 \Rightarrow S)
\end{align*}
\]

\( \psi_1 \) and \( \psi_2 \) follow from the minimal conditional dependence between \( X_1 \) and \( X_2 \) given \( X_3 \), \( \psi_3/\psi_4 \) from the minimal conditional independencies between \( X_1/X_2 \) and \( X_4 \) given \( X_3 \). Applying the causal rules (3.2), we can infer various new statements:

\[
\begin{align*}
(\psi_1 \land \psi_3) \vdash \gamma_1 & \quad \text{with} \quad \gamma_1 : (X_3 \Rightarrow X_4) \land (X_3 \Rightarrow S) \\
(\psi_2 \land \psi_4) \vdash \gamma_1 \\
(\psi_1 \land \psi_3) \vdash \gamma_2 & \quad \text{with} \quad \gamma_2 : (X_4 \Rightarrow X_1) \land (X_4 \Rightarrow S) \\
(\psi_2 \land \psi_4) \vdash \gamma_3 & \quad \text{with} \quad \gamma_3 : (X_4 \Rightarrow X_2) \land (X_4 \Rightarrow S)
\end{align*}
\]

The derivation of \( \gamma_1 \) is relatively straightforward; \( \gamma_2 \) and \( \gamma_3 \) are most easily proven by contradiction. Note further that two combinations here lead to the same statement. Given these new statements \( \gamma_1 \) through \( \gamma_3 \), we can then also infer

\[
\begin{align*}
\gamma_1 \land \gamma_2 & \vdash \theta_1 \quad \text{with} \quad \theta_1 : (X_4 \Rightarrow X_3) \land (X_4 \Rightarrow S) \\
\gamma_1 \land \gamma_3 & \vdash \theta_1
\end{align*}
\]

Figure 3.1 gives the logic tree for deriving the causal statement \( \theta_1 \). The question we would like to answer is: given probabilities of the base causal statements \( \psi_1 \) through \( \psi_4 \), what we can say about the probability of the inferred statement \( \theta_1 \)? Note that in this example, BCCD with \( K = 5 \) would already give \( \theta_1 \) as a base causal statement with a corresponding reliability.


Chapter 3. Computing lower and upper bounds on the probability of causal statements

Figure 3.1: (a) Example of a Y-structure. (b) Different levels of representing the inference of the statement $\theta$ that encodes an arrow between variable $X_4$ and variable $X_3$ in Figure (a).

3.2.2 Estimation of Probability Intervals

The native version of BCCD estimates reliabilities from inferred statements by taking the product when statements are combined with an AND (as if the underlying statements are independent) and taking the maximum when statements are combined with an OR (giving a conservative estimate). Empirically, this appears to work fine in practice. Here, we propose to give up on estimating reliabilities on inferred statements, but instead derive an algorithm to compute reliability intervals making use of the well-known Fréchet inequalities (Fréchet, 1935).

As should be clear from Figure 3.1, newly inferred statements are derived from a (potentially complicated) mixture of conjunctions (ANDs) and disjunctions (ORs) of the base statements. Suppose we have the conjunction $\psi_{\text{conjunct}} \leftarrow \psi_1 \land \psi_2 \land \ldots \land \psi_n$, then the Fréchet inequalities give

$$\max \left( 0, \sum_{i=1}^{n} P(\psi_i) - (n - 1) \right) \leq P(\psi_{\text{conjunct}}) \leq \min_{i} P(\psi_i),$$

with $P(\psi_i)$ the probability of the causal statement $\psi_i$. Similarly, applying Fréchet inequalities to a disjunction $\psi_{\text{disjunct}} \leftarrow \psi_1 \lor \psi_2 \lor \ldots \lor \psi_n$ gives

$$\max_{i} P(\psi_i) \leq P(\psi_{\text{disjunct}}) \leq \min \left( 1, \sum_{i=1}^{n} P(\psi_i) \right).$$

Using these inequalities, we can keep track of lower bounds and upper bounds on the probability of inferred causal statements, indicated by $P$ and $\bar{P}$, respectively. We will use shorthand notation $I(\psi) = [P(\psi), \bar{P}(\psi)]$ to refer to the probability interval for causal statement $\psi$. We will use lower case Greek letters $\psi, \gamma, \theta$ to refer to causal statements and upper case Greek letters $\Psi, \Gamma, \Theta$ for the
3.2. A Method for Computing Causal Reliability Intervals

corresponding formulae that have been used to derive the causal statements. We will write both \( \bar{P}(\Phi) \) and \( \bar{P}(\phi) \), the interpretation of which should be clear from the context.

Specifically, we have the following rules for combining intervals of two causal statements:

\[
\begin{align*}
\bar{P}(\psi_1 \land \psi_2) &= \max (0, P(\psi_1) + P(\psi_2) - 1), \\
\bar{P}(\psi_1 \lor \psi_2) &= \min (\bar{P}(\psi_1), \bar{P}(\psi_2)), \\
P(\psi_1 \lor \psi_2) &= \max (P(\psi_1), P(\psi_2)), \\
P(\psi_1 \land \psi_2) &= \min (1, P(\psi_1) + P(\psi_2)).
\end{align*}
\]  

(3.3)

We now propose two different algorithms: greedy and delayed. When two causal statements are combined to derive a new one, the greedy algorithm immediately applies the rules (3.3) to compute a new interval for the newly inferred statement. The delayed algorithm, on the other hand, delays the computation of the intervals as much as possible. It keeps track of the propositional formula that led to the causal statement of interest, attempts to simplify it, and only then computes the interval. We can use each of these algorithms as subroutines in the second part of the BCCD algorithm, explained in Section 3.2.1.

Example 2. We consider the logic tree of Figure 3.1. Suppose that the probabilities of the base causal statements are \( P(\psi_1) = 0.8, P(\psi_2) = 0.85, P(\psi_3) = 0.9, \) and \( P(\psi_4) = 0.95 \). The greedy algorithm sequentially applies (3.3) to yield \( I(\gamma_2) = [\max(0, P(\psi_1) + P(\psi_3) - 1), \min(P(\psi_1), P(\psi_3))] = [\max(0, 0.8 + 0.9 - 1), \min(0.8, 0.9)] = [0.7, 0.8], \) \( I(\gamma_3) = [0.8, 0.85], I(\gamma_1) = [0.8, 1], I(\gamma_1 \land \gamma_2) = [0.5, 0.8], I(\gamma_1 \land \gamma_3) = [0.6, 0.85], \) and, finally, \( I(\theta_1) = [0.6, 1] \). The delayed algorithm, on the other hand, first expresses each of the inferred causal statements as a formula in terms of the base statements \( \psi \) through \( \psi_4 \). For \( \gamma_1 \) and \( \theta_1 \), we arrive at

\[
\theta_1 \vdash \Theta_1, \gamma_1 \vdash \Gamma_1 \quad \text{with} \quad \Theta_1 \equiv \Gamma_1 = (\psi_1 \land \psi_3) \lor (\psi_2 \land \psi_4).
\]

(3.4)

Now applying the Fréchet inequalities, first on the conjunctions and then on the disjunction, we obtain \( I(\theta_1) = I(\gamma_1) = [0.8, 1] \): a tighter bound than for the greedy algorithm.

A causal statement combining more than two base causal statements can have various equivalent formulae, each suggesting a different ordering in the application of the Fréchet inequalities. In the above example, the formula (3.4) happens to be in disjunctive normal form (DNF): a disjunction of clauses, each of which is a conjunction of literals. Instead, we could also write it in conjunctive normal form (CNF), as a conjunction of disjunctive terms:

\[
\Theta_1 = (\psi_1 \lor \psi_2) \land (\psi_1 \lor \psi_4) \land (\psi_3 \lor \psi_2) \land (\psi_3 \lor \psi_4).
\]

(3.5)

Given this formula, we would first apply the Fréchet inequalities to the disjunctive clauses and only then to their conjunction. This would give \( I(\theta_1) = [0.65, 1] \).
Each formula is monotone, i.e., only contains positive statements. Note that here we treat \((Z \Rightarrow X) \land (Z \Rightarrow S)\) and \((Z \Rightarrow X) \lor (Z \Rightarrow S)\) as two separate (positive) causal statements, where one follows from the other because of the acyclicity condition. The minimal DNF and CNF representations corresponding to a monotone formula are unique (Goldsmith et al., 2005). Given a formula \(\Gamma\), we now consider two ways of computing bounds: following the natural ordering corresponding to its minimal DNF representation \(\Gamma_{\text{minDNF}}\) or according to its minimal CNF representation \(\Gamma_{\text{minCNF}}\). In the Appendix, we show that, when using Fréchet inequalities, the minimal DNF representation is better for computing the lower bound, i.e., \(\bar{P}(\Gamma_{\text{minDNF}}) \geq \bar{P}(\Gamma)\), whereas the minimal CNF representation is better for computing the upper bound, i.e., \(\bar{P}(\Gamma_{\text{minCNF}}) \leq \bar{P}(\Gamma)\).

This then suggests the following approach for the delayed algorithm. For each statement, we keep track of its minimal DNF and CNF. Whenever we combine two statements \(\gamma_1\) and \(\gamma_2\) with corresponding formulae \(\Gamma_1\) and \(\Gamma_2\) to derive a novel statement \(\theta_1\) using a conjunction, we combine the minimal CNFs of \(\Gamma_1\) and \(\Gamma_2\) into a CNF through \(\Theta_1 = (\Gamma_{\text{minCNF},1} \land \Gamma_{\text{minCNF},2})\), simplify that to its minimal CNF, and convert this to a (minimal) DNF using Quine’s algorithm (Quine, 1955). When it so happens that the novel statement \(\theta_1\) coincides with an earlier derived statement \(\theta_2\), we keep only one statement and replace \(\Theta_1\) and \(\Theta_2\) by their disjunction \((\Theta_1 \lor \Theta_2)\) simplified into its minimal DNF.

In practice, the total number of literals in the formulae may grow very fast. Since translating between CNF and DNF can produce expressions of exponential size, the delayed algorithm is practically infeasible. In practice, we therefore restrict the number of literals in the minimal DNF representation to a prespecified maximum \(M\). That is, if combining two causal statements \(\gamma_1\) and \(\gamma_2\) with corresponding formulae \(\Gamma_1\) and \(\Gamma_2\) leads to a new formula \(\Theta_{\text{minDNF}}\) with in total more than \(M\) literals, we choose to ignore how \(\gamma_1\) and \(\gamma_2\) were derived and switch to a greedy approximation at this level, cutting the tree and treating \(\gamma_1\) and \(\gamma_2\) as new base statements with their respective lower and upper bounds. In the experiments described in the next section, we set \(M = 13\).

## 3.3 Results

### 3.3.1 Simulated Data

Through a simulation study, we aim to investigate to what extent the type of algorithm (greedy and delayed) affects the tightness of the bounds. Furthermore, we will check whether any improvement in the bounds then also leads to an improvement in the accuracy of the causal statements derived.
We generated data with sample sizes 500, 1500, and 3000 from random graphs with linear interactions between 6, 9, and 12 Gaussian random variables and some other predefined properties (adapted from \cite{Melancon2000}). Each experiment is repeated 20 times. For the delayed algorithm, we set the maximum number of literals in any formula to $M = 13$, to obtain a compromise between accuracy and computational complexity. Higher values do not lead to significantly different results, but considerably increase the computational costs. We considered BCCD with two different values, $K = 3$ and $K = 5$ (the default), for the parameter $K$ that specifies the maximum number of variables used to infer the base causal statements. Since with higher $K$, BCCD will already find many causal statements without the need to explicitly combine statements using causal logic, we expect the difference between the greedy and the delayed algorithm to be more distinct for the smaller value of $K$. As in native BCCD, we process causal statements sequentially, going from the most to least reliable (in terms of the lower bound reliability) and ignore any causal statements with a (lower bound) reliability below 0.5.

We first focus on the lower bounds obtained by both variants for causal statements involving only pairs of variables (and possibly selection bias), i.e., statements of the form $Z \Rightarrow X$ (with and without potential selection bias on $Z$) and $Z \Leftrightarrow X$. We will refer to these as pairwise causal statements. Since for many such pairwise causal statements the greedy and delayed variants give the exact same lower bound, we only consider those cases where the lower bounds are indeed different. Because both variants ignore any causal statements with a lower bound reliability below 0.5, the greedy variant will typically end up with less pairwise causal statements than the delayed variant. When comparing the lower bound reliability for causal statements that are only inferred by the delayed variant, we set the lower bound for the greedy algorithm to 0.5.

The plots on the top of Figure 3.2 give the mean difference between the lower bounds inferred by both variants for different sample sizes, numbers of variables, and $K = 3$ (green, solid) or $K = 5$ (blue, dashed), averaged over 20 experiments. The errorbars give the 95% confidence interval of the mean. With $K = 3$, the delayed algorithm indeed improves the lower bound, in particular for larger sample sizes where base causal statements tend to have a higher reliability and more causal statements can be combined before the reliability drops below the threshold of 0.5. The size of the graph does not appear to have a serious effect. For $K = 5$, the improvement in the tightness of the lower bound is smaller: most causal statements are directly derived as base causal statements.

Having access to the ground truth, we can compare the inferred causal structures, in PAG form, with the true underlying PAG. Here, for each output PAG, we evaluate how many edges are correctly oriented by each of the algorithms.
Chapter 3. Computing lower and upper bounds on the probability of causal statements

Figure 3.2: Results on simulated data. Top row: mean difference between the lower bound reliabilities for pairwise causal statements computed using the delayed and the greedy variants as a function of sample size. Bottom row: difference in PAG accuracy (% of edge marks in PAG) for both variants as a function of sample size. Results for graphs with 6, 9, and 12 variables (from left to right) and $K = 3$ (green, solid) and $K = 5$ (blue, dashed). See the main text for further explanation.

For example if the model predicted the edge direction as $X_1 \rightarrow X_2$, while the ground model was $X_1 \leftrightarrow X_2$, we count that arrow from $X_1$ to $X_2$ was predicted correctly, but tail was predicted incorrectly and count one error in two causal statements. The plots on the bottom of Figure 3.2 display the mean difference between the PAG accuracy of the delayed and the greedy variant, over the same 20 experiments. Errorbars again give the standard error of the mean. It can be seen that, for larger sample sizes, the better bounds indeed appear to lead to a small improvement for the delayed over the greedy algorithm. In conclusion, the more expensive delayed algorithm leads to better bounds. The greedy algorithm is to be preferred when computational complexity is really an issue and then the greedy algorithm still leads to acceptable bounds and PAG accuracies.

3.3.2 Real-World Data

To illustrate the estimate of the lower and upper bound by the delayed algorithm ($K = 5$) on real-world data, we use the data described in (Hoogman et al., 2013, 2011) describing patients with attention deficit-hyperactivity disorder (ADHD). The study included 164 participants, 87 patients, and 77 control subjects from the Dutch chapter of the International Multicentre persistent ADHD Collaboration (IMpACT). The goal of the study was to investigate the connection between
candidate gene DAT1, brain functioning, and behavior characteristics associated with reward-related problems in ADHD. Two experiments were performed to learn these relationships. In the first experiment brain activity during the reward anticipation phase of the Monetary Incentive Delay (MID) task in a functional MRI paradigm was assessed. The MID task activates the ventral striatum, where DAT1 is most highly expressed. In the second experiment a delay discount task (Dom et al., 2006) was performed that aims to evaluate reward-related impulsivity.

To apply causal discovery using the BCCD algorithm, we selected ten variables from this data set. The first seven variables (disease status, smoking behavior, hyperactivity/impulsivity symptom score, attention-deficit symptoms score, medication status, presence of the DAT1 risk haplotype, ventral-striatal brain activation) were selected as part of the study that is described in detail in the next chapter. The extra three variables that were added based on a delay discount task are: reward-related impulsivity behavior; IQ level; education level. As prior information we incorporated the assumption that the DAT1 risk haplotype cannot be influenced by any other factor in the model, and that diagnosis is present downstream of symptoms, i.e., that symptoms cannot be caused by diagnosis.

The resulting causal graph is presented in Figure 3.3. This figure includes only edges with a reliability of a direct causal link higher than 50%. The edges inferred in the graph are in line with several literature studies. The link between DAT1 risk haplotype and ADHD symptoms is also found in (Gizer et al., 2009). The effect of the ADHD on brain functioning was shown in (Scheres et al., 2007). The association between ADHD and smoking was described in (Milberger et al., 1997). Correlation between ADHD and reward-related impulsivity was discussed in (Paloyelis et al., 2009). Thus, we can conclude that BCCD has inferred a reliable skeleton from the data.

BCCD was also able to infer the directions of some edges in Figure 3.3. Here we used a lower bound threshold of 30% to get a broader overview of possible edge directions. Combining causal statements, BCCD was able to infer the lower and upper bounds for two arrows and two tails in the graph. Other edge directions were directly inferred from the conditional independencies observed in the data. The lower and upper bounds for the edge directions suggest that there is strong evidence that ADHD status (patient/control) influences the medication status (reliability tail [0.78,1.0], arrow [0.89,0.99]). On the other hand there is vague evidence that the association between reward-related impulsivity and IQ goes from former to latter (reliability tail [0.5,1.0]). We compared these bounds with bounds obtained by the default BCCD algorithm that estimates the probability of a causal statement by taking the product of probabilities in case of a conjunction and maximum in case of a disjunction. Using the default
Figure 3.3: The causal graph representing causal relationships between variables for the ADHD data set. The graph represents a PAG, where edge directions are marked with "−" and "→" for invariant edge directions and with "○" for non-invariant edge directions. The reliability of an edge and its direction are depicted with a reliability score in the interval [0,1] near each edge, in the following format: "the arrow or tail on the left/edge/arrow or tail on the right". In case the annotation is vertical then top row is the reliability of the top arrow or tail and bottom row is the reliability of the bottom arrow or tail. Edge directions where the reliability of the arrow or tail is lower than 30% are marked with "-".

algorithm the probability of a tail between reward-related impulsivity and IQ was 0.56 and the probability of an arrow was 0.46. Thus, both probabilities were slightly higher than the lower bound obtained with the delayed algorithm. The probability of an arrow and a tail between ADHD status (patient/control) and the medication status obtained with the default BCCD algorithm coincided with the lower bound obtained with the delayed algorithm.

3.4 Conclusion and Discussion

In this chapter we provided a method to estimate lower and upper bounds for the reliability of a causal statement. Such bounds are valuable to convey the uncertainties involved in causal discovery to practitioners. They help to provide guidelines for setting up new (intervention) experiments to test causal hypotheses inferred from observational data.

We demonstrated how our approach can be integrated in the BCCD algorithm. However, any other constraint-based algorithm that infers causal state-
ments can be used potentially. We showed that a full, “delayed” version of our algorithm gives the best bound and PAG accuracies. However, it can be expensive for larger graphs, which is why we came up with a cheaper, “greedy” variant. In this chapter we considered Fréchet inequalities to estimate the lower and upper bounds of causal statements, but other approaches such as linear programming can also be used to compute the bounds. For example, using SAT solvers in some cases it is possible to provide more accurate lower and upper bound estimates than using Fréchet inequalities, as shown in (Hailperin, 1965).

3.5 Appendix

First we recall some well-known concepts from Boolean logic. Literals are variables and negated variables. A conjunction of literals is a term, sometimes represented as a set of literals. A disjunction of literals is a clause. Every Boolean function can be represented as a conjunction of clauses, referred to as conjunctive normal form (CNF), as well as a disjunction of terms (DNF), referred to as disjunctive normal form (DNF). A monotone Boolean function is one without any negated variables. A term $\phi$ subsumes a term $\psi$ iff $\phi \subset \psi$.

Lemma 3.1 (Lower bound). Given any monotone formula $\Gamma$ and its corresponding (unique) minimal DNF representation $\Gamma_{\text{minDNF}}$. When using Fréchet inequalities, the minimal DNF representation gives the best possible lower bound, i.e., $P(\Gamma_{\text{minDNF}}) \geq P(\Gamma)$.

Proof. The proof is by induction on the number $n$ of operators ($\lor, \land$) in $\Gamma$. The base case $n = 0$ clearly holds. If $n > 0$ then $\Gamma$ can be rewritten as a disjunction, $\Gamma = \Gamma_1 \lor \Gamma_2$ or a conjunction, $\Gamma = \Gamma_1 \land \Gamma_2$ of two formulae $\Gamma_1$ and $\Gamma_2$, with $(n_1, n_2) \leq n$. We assume that the lemma holds for $\Gamma_1$ and $\Gamma_2$, and prove that it then also holds for $\Gamma$. Let us first consider the disjunction, i.e., we suppose that $\Gamma = \Gamma_1 \lor \Gamma_2$. Then, by definition of how we apply the Fréchet inequalities,

$$P(\Gamma) = P(\Gamma_1 \lor \Gamma_2) = \max\{P(\Gamma_1), P(\Gamma_2)\} \leq \max\{P(\Gamma_{\text{minDNF},1}), P(\Gamma_{\text{minDNF},2})\},$$

where the last step follows from the induction assumption. Now,

$$P(\Gamma) \leq \max\{P(\Gamma_{\text{minDNF},1}), P(\Gamma_{\text{minDNF},2})\} = P(\Gamma_{\text{minDNF},1} \lor \Gamma_{\text{minDNF},2}) = P(\Gamma_{\text{minDNF}}).$$

Since all formulae are monotone, $\Gamma_{\text{minDNF}}$ is unique and can be obtained by removing terms from the disjunction of $\Gamma_{\text{minDNF},1}$ and $\Gamma_{\text{minDNF},2}$ such that it does not contain any subsumed terms (Quine, 1955). Clearly, removing subsumed terms
does not change the lower bound, which gives the last step in the disjunctive part of the proof. For the conjunction, we have

\[ P(\Gamma) = P(\Gamma_1 \land \Gamma_2) = \max (0, P(\Gamma_1) + P(\Gamma_2) - 1) \]

\[ \leq \max (0, P(\Gamma_{\min DNF,1}) + P(\Gamma_{\min DNF,2}) - 1) , \]

again with the first step by definition of how we compute a lower bound using the Fréchet inequalities and the last step from the induction assumption. Now, since for probability values \((a, b) \in [0, 1]\), we have \(a + b - 1 \leq \max(a, b)\), we get

\[ P(\Gamma) \leq \max (0, P(\Gamma_{\min DNF,1}) + P(\Gamma_{\min DNF,2}) - 1) \]

\[ \leq \max (P(\Gamma_{\min DNF,1}), P(\Gamma_{\min DNF,2})) = P(\Gamma_{\min DNF,1} \lor \Gamma_{\min DNF,2}) = P(\Gamma_{\min DNF}) . \]

**Lemma 3.2 (Upper bound).** Given any monotone formula \(\Gamma\) and its corresponding (unique) minimal CNF representation \(\Gamma_{\min CNF}\). When using Fréchet inequalities, the minimal CNF representation gives the best possible upper bound, i.e., \(\bar{P}(\Gamma_{\min CNF}) \leq \bar{P}(\Gamma)\).

**Proof.** We follow exactly the same reasoning as in the proof of Lemma 1, but consider the upper bound instead of the lower bound and minCNF instead of minDNF. For \(\Gamma = \Gamma_1 \land \Gamma_2\) we now have

\[ \bar{P}(\Gamma) = \bar{P}(\Gamma_1 \land \Gamma_2) = \min (\bar{P}(\Gamma_1), \bar{P}(\Gamma_2)) \geq \min (\bar{P}(\Gamma_{\min CNF,1}), \bar{P}(\Gamma_{\min CNF,2})) , \]

and

\[ \bar{P}(\Gamma) \geq \min (\bar{P}(\Gamma_{\min CNF,1}), \bar{P}(\Gamma_{\min CNF,2})) \]

\[ = \bar{P}(\Gamma_{\min CNF,1} \land \Gamma_{\min CNF,2}) = \bar{P}(\Gamma_{\min CNF}) . \]

For \(\Gamma = \Gamma_1 \lor \Gamma_2\) we have

\[ P(\Gamma) = P(\Gamma_1 \lor \Gamma_2) = \min (1, P(\Gamma_1) + P(\Gamma_2)) \]

\[ \geq \min (1, P(\Gamma_{\min CNF,1}) + P(\Gamma_{\min CNF,2})) , \]

and

\[ \bar{P}(\Gamma) \geq \min (1, \bar{P}(\Gamma_{\min CNF,1}) + \bar{P}(\Gamma_{\min CNF,2})) \geq \min (\bar{P}(\Gamma_{\min CNF,1}), \bar{P}(\Gamma_{\min CNF,2})) \]

\[ = \bar{P}(\Gamma_{\min CNF,1} \land \Gamma_{\min CNF,2}) = \bar{P}(\Gamma_{\min CNF}) . \]

---

56
In this chapter we focus on the application of the BCCD algorithm discussed in Chapters 2 and 3 to the real-world data about Attention-deficit/hyperactivity disorder (ADHD). This disorder is highly heritable, however, little is known about the genetic mechanisms that cause it. One of the candidate genes for ADHD is DAT1, encoding the dopamine transporter. In an attempt to clarify its mode of action, brain activity was accessed during the reward anticipation phase of the Monetary Incentive Delay (MID) task in a functional MRI paradigm in 87 adult participants with ADHD and 77 controls (average age 36.5 years). The MID task activates the ventral striatum, where DAT1 is most highly expressed. A previous analysis based on standard statistical techniques did not show any significant dependencies between a variant in the DAT1 gene and brain activation. In this chapter we used an alternative method for analyzing the data, causal modeling using BCCD algorithm (see Chapter 2). The main advantages of BCCD in comparison to other methods is it’s ability to find direct and indirect dependencies between variables, determine the strength of the dependencies, and provide a graphical visualization to interpret the results. BCCD confirmed that there is no evidence of a direct link between DAT1 genetic variability and brain activation, but suggested an indirect link mediated through inattention symptoms and diagnostic status of ADHD. Our finding of an indirect link of DAT1 with striatal activity during reward anticipa-

This chapter is based on Sokolova et al. (2015b), “Causal discovery in an adult ADHD data set suggests indirect link between DAT1 genetic variants and striatal brain activation during reward processing”, published in the American Journal of Medical Genetics Part B: Neuropsychiatric Genetics.
Indirect link between DAT1 genetic variants and striatal brain activation during reward processing might explain existing discrepancies in the current literature. Further experiments should confirm this hypothesis.

4.1 Introduction

Attention-deficit/hyperactivity disorder (ADHD) is an impairing and highly heritable disorder affecting both children and adults. It is characterized by two types of symptoms, hyperactivity/impulsivity and inattention, which can occur separately or combined. Studies show that individuals with ADHD have altered cognitive functioning in several domains, among which is reward processing. For example, people with ADHD tend to choose small immediate rewards instead of larger delayed rewards more frequently than healthy controls (Luman et al., 2005). Also on the neural level, there is evidence for altered reward processing. A key area in the brain involved in reward processing is the striatum, and many studies using magnetic resonance imaging (MRI) have shown hypoactivation in this part of the brain in individuals with ADHD. A recent meta-analysis showed a significant, moderate effect size (Cohen’s d = 0.48–0.58) of ventral–striatal (VS) hyporesponsiveness in individuals with ADHD (Plichta and Scheres, 2014).

The neurotransmitter system thought to be most strongly involved in reward processing in the brain is dopamine. Several dopamine-related genes have been associated with ADHD (Franke et al., 2012), among which the gene encoding the dopamine transporter, DAT1 (official name SLC6A3) (Franke et al., 2010; Gizer et al., 2009). Genetic variation of the DAT1 gene may lead to interindividual variation in the availability of dopamine transporters and, as a result, in the level of dopamine available for signaling (Faraone et al., 2014; Shumay et al., 2011). The 10-6 haplotype of two DAT1 variable number of tandem repeat (VNTR) polymorphisms, one located in the 3’ untranslated region (UTR) and the other in intron 8, has been shown to increase risk for ADHD in childhood (Asherson et al., 2007; Brookes et al., 2006). A different haplotype of the same two VNTRs, the 9-6 haplotype, was found associated with ADHD in adults (Franke et al., 2008, 2010).

Several studies have been performed in an attempt to clarify whether - and if so, how - DAT1 affects reward processing. The results of those studies have been inconsistent. Three studies in healthy adults found lower reward-related striatal activation during reward anticipation to be associated with homozygosity for the 10-repeat allele of the 3’UTR VNTR compared to carriersonhip of the 9-repeat allele (Aarts et al., 2010; Dreher et al., 2009; Forbes et al., 2009), two other studies did not find an effect of DAT1 genotype (Hahn et al., 2011; Nikolova et al., 2011). In children with ADHD, one study found lower striatal activation to be associated
4.2. Materials and Methods

4.2.1 Participants

The study included 164 participants, 87 patients and 77 control subjects from the Dutch chapter of the International Multicentre persistent ADHD CollaboraTion (IMpACT) (Franke et al., 2010). Patients and controls represented two age-, gender-, and IQ-comparable groups. All subjects participated in cognitive testing and neuroimaging. Twenty-seven of the patients who underwent the with the homozygous 10-repeat compared to 9-repeat carrierrship (Durston et al., 2008), whereas another study found the opposite (Bedard et al., 2010).

To help resolve inconsistencies between previous studies, Hoogman et al. (2013) assessed brain activity in a sample of adults with ADHD and healthy comparison subjects during a Monetary Incentive Delay (MID) task. This task is known to activate the ventral striatum (Hermans et al., 2010; Knutson et al., 2001), where DAT1 is most highly expressed. Analyzing 87 patients and 77 controls, the standard statistical methods (F-test, t-test, $X^2$ test) used in this analysis (a) showed higher prevalence of the 9-6 DAT1 haplotype in patients compared to the healthy control group ($X^2=10.04$, $p=0.002$), (Hoogman et al., 2013), and (b) confirmed previous findings that individuals with ADHD have lower task-related striatal activation compared to healthy subjects ($t(162) =-2.32$, $p=0.02$). The study, however, did not reveal a significant effect of the DAT1 haplotype on striatal activation during reward anticipation ($F(3,158) =0.24$, $p=0.63$).

In the current study we used an alternative method to analyze the data reported in Hoogman et al. (2013), involving causal modeling. Bayesian Constraint-based Causal Discovery (BCCD) algorithm (Claassen and Heskes, 2012a) described in detail in Chapter 2 applied in this study learns a causal model from the observed data. This method focuses on the exploratory analysis of the data, suggesting probable causal dependencies and providing novel hypotheses for further testing. The causal modeling approach has several advantages over standard analysis techniques. First, it provides an opportunity to learn the causes and effects from data. Second, it detects whether the dependency between variables is direct or mediated through other variables. Third, it can visualize the results in the form of a graph that makes the interpretation of the results easier. The goal of this study was to make an exploratory analysis of the data describing MID task using the BCCD algorithm to build a model that can explain the inconsistencies between previous studies. Thus, no a priori hypotheses about the effect of gene on brain functioning were made at the beginning of the study.
Chapter 4. Indirect link between DAT1 genetic variants and striatal brain activation during reward processing

tests were medication-naive. The rest used medication and had to withdraw it from 24 hours before the experiments. Patients were also asked to refrain from smoking and drinking coffee before and during testing. Diagnostic work-up of the patients and controls has been described elsewhere (Hoogman et al., 2013). Shortly, patients were included if they met DSM-IV-TR criteria for ADHD in childhood as well as adulthood. Participants were assessed using the Diagnostic Interview for Adult ADHD (DIVA) (Kooij, 2010) to confirm clinical diagnosis. In addition, a quantitative measure of clinical symptoms was obtained using the ADHD-DSM-IV Self Rating scale (Kooij et al., 2005). Additional measures included IQ (using 2 subtests of the Wechsler Adult Intelligence Scale-III), smoking behavior (self-report), and medication status (self-report).

Exclusion criteria for participants were psychosis, addiction in the last 6 months, current major depression (assessed with SCID-I), full-scale IQ estimate less than 70 (Wechsler Adult Intelligence Scale-III), neurological disorders, sensorimotor handicaps, non-Caucasian ethnicity, and medication use other than psychostimulants or atomoxetine. Additional exclusion criteria for comparison subjects were a current or past neurological or psychiatric disorder according to SCID-I.

4.2.2 Experiment description

In order to evaluate striatal activation in response to the reward stimulus, Hoogman et al. (2013) performed a set of experiments involving the MID task (Knutson et al., 2001). While performing the task, participants had to press a button as soon as possible, when seeing a target on the screen. If doing so in time, a reward could be earned. Prior to the target screen a reward cue screen was shown, which indicated if the reward could be obtained or not. After each response, the outcome was displayed. The order of the stimuli was as follows: cue (reward or no reward cue) with a duration of 4 to 9 seconds, followed by a target which was response active. After this a feedback screen was shown for 500 msec followed by a blank screen for 500 msec, after which another cue was presented. This sequence of stimuli was shown 22 to 25 times for reward cues and 22 to 25 times for no-reward cues. The response window was individually adjusted in order to balance hits and misses. After a hit, 20 msec was subtracted from the maximum response time and 10 msec was added after a miss. This procedure resulted in comparable hit rates (35% for no-reward and 40% in the reward condition). The participants could gain 1 Euro in the reward condition and no money during the non-reward condition, if they responded between 270 and 500 ms after target onset. Whole brain imaging was performed with a 1.5 Tesla MR scanner (Magnetom Avanto, Siemens Medical Systems, Erlangen, Germany) and
a standard 8 channel head coil.

4.2.3 Data description

To apply causal discovery using the BCCD algorithm, we selected the following seven variables measured during the study that may influence striatal activation:

- Disease status (binary: patient/control), describing whether a participant was a patient or a control.

- Smoking behavior (binary: smoking/non-smoking), describing whether a participant was a smoker or not. This variable was chosen, since smoking influences the dopamine transmission in the brain and can therefore influence the results of the fMRI experiments (Brody, 2006).

- Medication status (binary: naive/not naive), describing whether a patient had ever used medication to treat his/her ADHD.

- Hyperactivity/impulsivity symptom score (discrete variable from 0 to 9), describing the self-reported current presence of the DSM-IV hyperactivity/impulsivity symptoms.

- Attention-deficit symptoms score (discrete variable from 0 to 9), describing the self-reported current presence of the DSM-IV inattention symptoms.

- Presence of the DAT1 haplotype (binary: present /absent), describing whether a participant carried at least one copy of the 9-6 DAT1 haplotype or not.

- Ventral-striatal brain activation during the MID task in the functional (f)MRI experiment (continuous variable, detailed explanation below), describing the level of activation of the ventral striatum during reward anticipation.

4.2.4 Brain data preprocessing

Data analysis of the fMRI images showed that in the contrast of brain activations during rewarded and non-rewarded trials, some subjects showed no increased brain activation in rewarded trials, whereas some had more than five brain regions activated. In order to correct for baseline activation we added an extra step to the normal processing pipeline of the fMRI data as described in Hoogman et al. (2013). For this, we picked a reference region, in which we did not expect any influence of reward/non-reward cue. In our case, we chose white
matter. Then we applied the standard procedure to find brain activation in the ventral-striatal region of interest (ROI) and the white matter ROI. The first step was to extract information about neural signals from the data obtained during the fMRI experiments, which also contains random noise and nuisance components. For this, a General Linear Model (GLM) was derived for the voxels that corresponded to the ROIs (Lindquist, 2008). In a second step, the neural signal was divided into two groups based on the reward and non-reward cues, in order to find the difference in neural signal between the groups. Then the signal in the ROIs was averaged for each subject, and the difference between groups was computed. Before doing the latter, we compared the brain activation between the striatal ROI and the reference ROI, using a Student’s t-test. The reference region thus became a modified baseline for the neural signal. The result of the Student’s t-test was used as a new estimate for the brain activation of interest. The idea of the described technique is similar in spirit to a widely used method for baseline correction of resting state MRI images using white matter (Grol et al., 2007; Majdandzic et al., 2007; Verhagen, 2012).

Using the new brain activation measure we estimated the difference in striatal activation between patients and controls. The new striatal activation measure with baseline correction better captured the dependencies between patients and controls (t(162) = -2.69, p=0.008) than the standard measure (t(162)=-2.32, p=0.02), probably because using a reference region for baseline adjustment reduced the noise in the data and improved the accuracy of estimating striatal activation.

4.2.5 Causal Modeling

One way to represent causal models is through structural equation modeling (SEM). This is a widely used statistical technique for testing and estimating causal relationships in the field of medical research (Beran and Violato, 2010). Commonly used practice in medical research is to define the structure of a SEM manually (Neale and Schmitt, 2005). An alternative approach is to learn the structure of the SEM automatically, using causal discovery algorithms (Pearl, 2000).

There are two main approaches to learn the structure of a SEM automatically: score-based and constraint-based (Daly et al., 2011). An advantage of the score-based approach is that it provides a measure of reliability of inferred causal relations. This makes the interpretation of the results easier and prevents incorrect categorical decisions (Heckerman et al., 1999). A major drawback of the approach is that it relies on the causal sufficiency assumption which means that the algorithm cannot detect common confounders of the observed variables.
An advantage of the constraint-based approach is that it does not have to rely on the causal sufficiency assumption, and, as a result, can detect common causes of the observed variables (Spirtes et al., 2000). A disadvantage of this approach is that its output is not always reliable. A standard approach makes use of independence tests, making the results for borderline independencies/dependencies incorrect sometimes (Spirtes et al., 2000). The outcome of learning a network can be sensitive to such errors.

The method that we used to learn the structure of the SEM was developed by (Claassen and Heskes, 2012a) and is called Bayesian Constraint-based Causal Discovery (BCCD). This method aims to combine the strength of constraint-based and score-based approaches. This method is able to detect common causes of the observed variables similar to constraint-based approaches and provides a reliability measure of the inferred relationship like the score-based approach. This reliability measure gives a conservative estimate of the probability of a causal relation.

A recently extended version of BCCD can handle data that contains a mixture of discrete and continuous variables and does not require discretization that can lead to loss of information as described in Chapter 2. BCCD works with directed acyclic graphs that can contain latent variables. These graphs are called maximal ancestral graphs (MAG). All MAGs that represent the same set of conditional independencies form an equivalence class. The equivalence class for MAGs is called a partial ancestral graph (PAG). Edge directions in a PAG are marked with "-" and ">" if the direction is the same for all graphs belonging to the PAG and with "○ " otherwise. The BCCD algorithm produces PAGs as an output.

### 4.3 Results

Demographics of the study sample previously used in Hoogman et al. (2013) are shown in Table 6.1. Patients and controls represented age-, gender-, and IQ-comparable groups (p>0.22). We applied the BCCD algorithm to this data set. As prior information we incorporated the assumption that DAT1 genotype cannot be influenced by any other factor in the model, since chronologically a gene is the first factor present in the lifespan, and that diagnosis is present downstream of symptoms, i.e., that symptoms cannot be caused by diagnosis.

Running the BCCD algorithm provided three tables (Tables 4.2, 4.3 and 4.4). Table 4.2 presents the reliability of the causal statement: “A causes B”, both for direct and indirect causal effects. If there is an edge between A and B, it has a tail from A to B and A causes B in the PAG. For example, variable “Patient/Control” caused variable “Medication” with reliability 89%. Table 4.3 represents the reliability of the causal statement: “A does not cause B”, both for
direct and indirect causal effects. If there is an edge between A and B, it has an arrow head from B to A. For example, the variable “Smoking” does not cause variable “Patient/Control” with reliability 66%. The prior knowledge used in the model, i.e. that DAT1 haplotype is not caused by other variables and that diagnosis is present downstream of symptoms, is represented in Table 4.3 in cells with a reliability of 100%.

Table 4.4 provides the reliability of the statement that a direct causal link
Table 4.3: The reliability of the logical statement “A does not cause B”, where A is represented in rows and B in columns. The estimate is provided for logical statements with reliability of 50% or higher.

<table>
<thead>
<tr>
<th>Factor A</th>
<th>Striatal activation</th>
<th>Smoking</th>
<th>Hyperactivity</th>
<th>Inattention</th>
<th>Patient/Control</th>
<th>Medication</th>
<th>DAT1 haplotype</th>
</tr>
</thead>
<tbody>
<tr>
<td>Striatal activation</td>
<td>-</td>
<td>-</td>
<td>53%</td>
<td>53%</td>
<td>53%</td>
<td>-</td>
<td>100%</td>
</tr>
<tr>
<td>Smoking</td>
<td>-</td>
<td>-</td>
<td>63%</td>
<td>63%</td>
<td>63%</td>
<td>-</td>
<td>100%</td>
</tr>
<tr>
<td>Hyperactivity</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>100%</td>
</tr>
<tr>
<td>Inattention</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>100%</td>
</tr>
<tr>
<td>Patient/Control</td>
<td>-</td>
<td>-</td>
<td>100%</td>
<td>100%</td>
<td>-</td>
<td>-</td>
<td>100%</td>
</tr>
<tr>
<td>Medication</td>
<td>-</td>
<td>-</td>
<td>90%</td>
<td>90%</td>
<td>90%</td>
<td>-</td>
<td>100%</td>
</tr>
<tr>
<td>DAT1 haplotype</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>

Table 4.4: Reliability of direct links between two variables.

<table>
<thead>
<tr>
<th>Factor A</th>
<th>Striatal activation</th>
<th>Smoking</th>
<th>Hyperactivity</th>
<th>Inattention</th>
<th>Patient/Control</th>
<th>Medication</th>
<th>DAT1 haplotype</th>
</tr>
</thead>
<tbody>
<tr>
<td>Striatal activation</td>
<td>-</td>
<td>16%</td>
<td>17%</td>
<td>24%</td>
<td>51%</td>
<td>26%</td>
<td>23%</td>
</tr>
<tr>
<td>Smoking</td>
<td>-</td>
<td>-</td>
<td>14%</td>
<td>29%</td>
<td>65%</td>
<td>15%</td>
<td>19%</td>
</tr>
<tr>
<td>Hyperactivity</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>&gt;99%</td>
<td>&gt;99%</td>
<td>19%</td>
<td>15%</td>
</tr>
<tr>
<td>Inattention</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>&gt;99%</td>
<td>10%</td>
<td>96%</td>
</tr>
<tr>
<td>Patient/Control</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>&gt;99%</td>
<td>-</td>
<td>7%</td>
</tr>
<tr>
<td>Medication</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>8%</td>
</tr>
<tr>
<td>DAT1 haplotype</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>

exists between two variables. The difference between Tables 4.2 and 4.3 compared to Table 4.4 is that Tables 4.2 and 4.3 give an estimate for the direction of the causal effect, whereas Table 4.4 provides estimates for the presence of the path between two variables. Moreover, Tables 4.2 and 4.3 show the reliability of the direction of both direct and indirect causal paths, whereas Table 4.4 gives a reliability estimate for a direct causal path between two variables only.

By combining Tables 4.2, 4.3 and 4.4 we constructed a causal network representing causal relationships between variables, which is presented in Figure 4.1. For visualization purposes, this network only includes the edges with a reliability of a direct causal link higher than 50%. The resulting network structure matched some of our expectations: symptoms caused diagnosis, and the presence of an ADHD diagnosis influenced smoking behavior, prescription of medical treatment, and level of striatal activation during the MID task. Moreover, the graph showed a direct link between the DAT1 haplotype and inattention symptoms, but not between the DAT1 haplotype and hyperactivity/impulsivity symptoms. These findings are in line with the results obtained by Hoogman et al. (2013).
As apparent from the network in Figure 4.1, the causal path from the DAT1 risk haplotype to brain activation is indirect, and mediated through other variables. This causal path was not detected in the study of Hoogman et al. (2013). Another causal path that had not been detected by the former study is the direct path between inattention and hyperactivity/impulsivity. Assuming that there is no common cause between the DAT1 haplotype and inattention symptoms, we measured the strength of the causal effect using Cohen’s d and odds ratio tests, depending on the variables of interest (Cohen, 1988; Nakagawa and Cuthill, 2007).

The causal effect of the DAT1 risk haplotype on the inattention symptoms has a large effect size (Cohen’s d 0.8, CI= [0.41, 1.20]). The link from inattention symptoms to patient/control status has a large odds ratio of 5.5 (CI= [2.87, 10.43]), while the link from patient/control status to striatal brain activation has a medium effect size (Cohen’s d 0.4, CI= [0.11, 0.73]). As a results, the direct effect of the DAT1 risk haplotype on brain activation was small and non-significant.
4.4 Discussion

In the current study we proposed an alternative method of data analysis for functional MRI, and behavioral and genetic data. The standard methods for such data analysis involve statistical tests that tell whether the difference in means between two populations is statistically significant. These methods are easy to use, but are restricted in the types of questions they can answer. They are focused only on the presence of the dependency between observed variables (the ones that are directly measured) and fail to determine the direction of this dependency. Moreover, if the dependency between two observed variables is mediated through a third observed variable, the standard methods would not detect it, but only indicate that all three variables are correlated.

The method proposed in this chapter has several advantages over the standard statistical techniques. It allows deeper insights into the data by building a complete model, instead of considering only pairwise dependencies, and by distinguishing between direct and indirect causal effects. Here, we applied the

(Cohen’s d 0.14, CI= [-0.24, 0.52]). If we leave out the variable patient/control, then the individual link between the inattention and hyperactivity and brain activation does not pass the threshold. We interpret the variable patient/control as a summary variable that combines the influence of hyperactivity and inattention and possibly other endophenotypic variables on the other variables in the model such as smoking, striatal activation and medication.

The causal path from diagnostic status to striatal activation appears to contradict models assuming that altered brain functioning is a cause of ADHD instead of it being a consequence, like the endophenotypic model proposed by Franke et al. (2009). If we enforce such an assumption, i.e., if we add the constraint that there cannot be a causal path from diagnosis to striatal activation, the causal path from the DAT1 haplotype to striatal activation disappears. Instead, in order to account for the observed correlation between diagnosis and striatal activation, our analysis yields either a causal link from striatal activation to diagnostic status, or a common cause (e.g., a comorbid disorder) that is associated with both ADHD and striatal activation. To estimate the probability of the two alternative networks with and without endophenotypic assumption, we calculated the Bayesian Information Criterion (Schwarz, 1978) and estimated the posterior odd ratios. The model with the endophenotypic assumption scored a factor of two worse (posterior odd ratio= 2.2) than the model without this assumption described in the results. However, it should be mentioned that this latter does not provide particularly strong evidence against the endophenotypic assumption.
BCCD approach to an existing data set of adult patients with ADHD and healthy controls, which had earlier been analyzed and published using standard methods of analysis (Hoogman et al., 2013).

In line with the earlier analysis results, our approach detected a direct link between the presence of ADHD and the level of brain activation during the MID test and a link between inattention and DAT1; neither the earlier nor the new method provided evidence for a direct link between DAT1 variation and brain activation during reward anticipation. The network built using the BCCD algorithm revealed additional causal paths that had not been detected by the earlier analysis. BCCD suggested that an indirect path exists between the DAT1 risk haplotype and striatal activation that is mediated through other variables, such as inattention. That might be explained by the idea that the processing of the reward or non-reward cues requires attention skills. The effect of the DAT1 risk haplotype on brain activation was small and did not reach formal statistical significance, thus being hard to detect using standard bivariate statistical techniques.

The existence of an indirect causal path from the DAT1 haplotype to striatal activation might explain existing discrepancies between findings in the literature. Studies performed earlier (Aarts et al., 2010; Dreher et al., 2009; Forbes et al., 2009) found an effect of the DAT1 haplotype on reward-related striatal activation, whereas others (Hahn et al., 2011; Nikolova et al., 2011) did not find this effect. The inferred indirect path shows that striatal activation depends on both inattention and hyperactivity, whereas the DAT1 haplotype appears to have a direct influence only on inattention. As a result, the effect of hyperactivity/impulsivity symptoms may blur the effect of DAT1 through inattention symptoms on striatal activation, which hence may appear statistically significant in one cohort, but not in another.

Our analysis further suggests that there is no direct causal link from DAT1 to hyperactivity/impulsivity: the observed correlation between the DAT1 haplotype and hyperactivity/impulsivity can be fully explained by the effect of DAT1 on inattention and an effect of inattention on hyperactivity/impulsivity. The latter is of interest in itself as well, as it might suggest that hyperactivity/impulsivity occurs downstream of inattention. This finding deserves further study, as it could have implications for ADHD treatment – focusing on treating inattention would then also reduce hyperactivity/impulsivity symptoms.

The model provided by our data challenges current ideas of brain imaging measures as endophenotypes for ADHD. Our data favored a model with an indirect causal path from the DAT1 haplotype to striatal activation mediated by symptoms and diagnosis, although, based on this first analysis, we do not claim to have strong evidence against the endophenotypic assumption. It is likely that
developmental brain alterations contribute to disease risk, and may be further enhanced or modified by disease symptoms potentially reversing the direction of the causal links. Recent findings by Cortese et al. (2013), for example, have suggested that brain alterations (in this case white matter alterations) cannot be linked to disease outcome in ADHD. Similarly, a recent treatment study suggests that improvements at the cognitive level upon drug treatment of ADHD do not correlate strongly with clinical improvement (Coghill et al., 2014). The general sparseness of papers describing convincing correlations of neural or cognitive findings with behavioral/clinical data suggests that such links are not readily detectable. Additional research is clearly necessary to investigate the link between brain imaging phenotypes and disease.

The strength of this study is the application of a novel causal discovery method for data analysis. This method considers all variables together, infers both direct and indirect dependencies between variables, provides a reliability measure for each edge in the network and is able to detect latent common causes.

The limitation of our causal discovery method is that it is an exploratory analysis – it provides new hypotheses that need to be tested using other methods and needs an independent replication. To verify this hypothesis experiments or additional data is required. Another limitation is the exclusion criteria during the selection of participants leading to the analysis of the data representing a subsample of the ADHD population, rather than the data representing the ADHD population seen in the daily practice. A potential limitation of our work is the fact that we could not assess the effect of medication duration, which could affect striatal activation (Schulz et al., 2012). In our analysis of the effects of medication, we pooled different stimulants, and given the numbers, we could not evaluate potential differential effects of medication. However, since we did not observe any significant differences between subjects using medication and those that were medication naïve, it is unlikely that the effects of such treatments have biased our results.

In conclusion, application of the BCCD algorithm confirmed that there is no statistical evidence for a direct link between DAT1 and ventral striatum activation during the MID task, but suggests that there is an indirect link mediated through inattention symptoms and diagnostic status. This finding might explain the inconsistency of results described in literature.
In this chapter we discuss the application of the causal modeling to different data sets about ADHD in order to better understand the correlation between ADHD symptom domains. Although numerous factor analytic studies consistently support a distinction between two symptom domains of ADHD, inattention and hyperactivity/impulsivity, it is not clear what drives strong correlation between them. To address this issue we used causal modeling on three independent data sets containing information about symptom scores and gender or a genetic risk haplotype. As a result we found strong statistical evidence for the same pattern: the clear dependence between hyperactivity/impulsivity symptom level and an established genetic factor (either gender or risk haplotype) vanishes when one conditions upon inattention symptom level. Relying on few reasonable assumptions, e.g., that phenotypes do not cause genotypes, an inferred causal model contains a causal path from inattention to hyperactivity/impulsivity. The robust dependency cancellation observed in three different data sets suggests that inattention is a driving factor for hyperactivity/impulsivity. This causal hypothesis can be further validated in intervention studies. Our model suggests that interventions that affect inattention will also have an effect on the level of hyperactivity/impulsivity. On the other hand, interventions that affect hyperactivity/impulsivity would not change the level of inattention. This causal model may explain earlier findings on heritable factors causing ADHD reported in the study of twins with learning difficulties.

This chapter is based on Sokolova et al. (2016a), “Statistical evidence suggests that inattention drives hyperactivity/impulsivity in attention deficit-hyperactivity disorder”, published in PLoS One.
5.1 Introduction

5.1.1 Problem description

Attention-deficit/hyperactivity disorder (ADHD) is a common and highly heritable neurodevelopmental disorder that affects about 5-6% of children worldwide (Polanczyk et al., 2007, 2014). ADHD persists into adulthood in about 30-50% of the childhood cases, depending on definition of remission (Faraone et al., 2006), and prevalence of ADHD in adults is estimated between 2.5-4.9% (Simon et al., 2009). In pediatric populations, ADHD is about 2-3 times more common in boys than girls (Bauermeister et al., 2007), but gender balance is rather equal in adult populations (Kooij et al., 2005). The genetics of ADHD is complex (Lahey et al., 2011) and several candidate genes have been associated with ADHD in meta-analyses, among which the dopamine transporter gene SLC6A3/DAT1 (Gizer et al., 2009) and dopamine D4 receptor gene DRD4 (Lichter et al., 1993). Genetic variation of the DAT1 gene may affect the functioning of the dopamine transporter caused by individual variation in regulating levels of dopamine (Shumay et al., 2011; Faraone et al., 2014). This alters baseline dopamine tone; which is utilized therapeutically by drugs such as methylphenidate that block the dopamine transporter involved in the recycling of dopamine into neurons. The DAT1 gene has a differential risk haplotype (formed by a variable number of tandem repeat (VNTR) polymorphisms in the 3′ UTR and in intron 8) associated with childhood ADHD (10R/6R) and adult ADHD (9R/6R) (Franke et al., 2008, 2010). Similarly, polymorphism in the 7 repeat allele of the DRD4 gene (which is expressed on neuronal dendrites) confers reduced intracellular cAMP signalling following binding of dopamine to dopamine D4 receptors. As such, increased expression of these VNTR polymorphisms in DAT1 or DRD4 increases the degree of genetic risk associated with ADHD symptoms. Furthermore, both DAT1 knockout and DRD4 knockout transgenic mice demonstrate face validity with documented increases in hyperactivity and impulsivity (Kooij and Glennon, 2007) and reduced behavioral inhibition (Avale et al., 2004).

As evident from its name, ADHD is characterized by inappropriate and pervasive levels of inattention and/or hyperactivity and impulsivity. Exploratory and confirmatory factor analyses of the core ADHD symptoms defined in the DSM system and assessed by parents and teachers, as well as self-report ratings in adolescents and adults consistently support a distinction between two symptom dimensions: inattention and hyperactivity/impulsivity (see (Willcutt et al., 2012) for a review). Inattention and hyperactivity/impulsivity both show high internal consistency and are moderately to strongly correlated (correlation coefficient between .63 and .75), indicating that they constitute separable but
substantially correlated dimensions (Willcutt et al., 2012). Inattention is more strongly related to internalizing problems of anxiety and depression and to academic underachievement. In contrast, hyperactivity/impulsivity is linked to peer rejection and externalizing behavioral problems such as oppositional defiant and antisocial behavior (Willcutt et al., 2012). The cause of the strong correlation between the two symptom dimensions of ADHD inattention and hyperactivity/impulsivity is yet unclear. Are these two dimensions two sides of the same coin, i.e., the consequence of a (possibly unknown) common cause, or could it be that one dimension drives the other? This question is relevant to the current literature: some studies assume a bi-factor model to explain the correlation (Martel et al., 2010), others propose a driving effect of inattention on hyperactivity based on the analysis of twin studies (Willcutt et al., 2000).

### 5.1.2 Causal discovery from observational data

The standard approach to establish causal relationships is through experimental manipulation or intervention. For example, in order to establish a causal effect of inattention upon hyperactivity/impulsivity, one would need to apply an intervention that only acts upon inattention and then measure its effect on hyperactivity/impulsivity. When analyzing the results of these experiments the Bradford Hill criteria for causation should be taken into account (Hill, 1965). These criteria specify the conditions necessary to provide evidence of causal relationships. Although in theory such an intervention, e.g., through a well-designed therapy or some novel highly specific medication, could be attainable, we are not aware of any such attempts or studies in the current literature.

That being the case, the emerging field of causal discovery from observational data may provide a powerful alternative (Pearl, 2000; Spirtes, 2010). In apparent contradiction with the good old adagio “correlation does not imply causation”, theoretical and experimental studies have shown that, under certain reasonable assumptions, it is possible to learn cause-effect relationships from purely observational data. The key insight is that, where a single number such as a mere correlation indeed cannot reveal anything about causal direction, other, more subtle characteristics may contain important directional information. Just considering pairs of variables, these can be found in higher-order moments (Mooij et al., 2014). In higher-dimensional systems, the seminal work of Turing award winner Judea Pearl (Pearl, 2000) and others revealed the close connection between causal relationships and conditional independencies. Since then, causal discovery algorithms have successfully been applied in various domains, and slowly find their way into the biomedical sciences (Schadt et al., 2005; Chen et al., 2007; Maathuis et al., 2010; Schmidberger et al., 2011). To the best of our
knowledge, the current study is the first to describe an application of causal discovery for the analysis of observational clinical data.

Intuitively, two variables $Z$ and $Y$ are conditionally independent given $X$ if, once the value of variable $X$ is known, the value of $Z$ does not add any additional information about $Y$. For example, in the context of children with ADHD, we can call gender and hyperactivity/impulsivity conditionally independent given inattention, if knowing whether a subject is a boy or a girl does not help to better estimate the hyperactivity/impulsivity symptom score, once we already know the child’s inattention symptom score. In this chapter we investigate whether such conditional independencies can be derived from observational data.

Most causal discovery algorithms start by assuming that real-world events are governed by specific, yet unknown causal mechanisms. Given a particular causal model, one can in principle read off the conditional dependencies and independencies one should then find in observational data. Reasoning backwards, given particular observed conditional dependencies and independencies in observational data, one may be able to infer causal relations that any causal model should have to be consistent with the observed statistical patterns.

It is exactly this kind of inverse reasoning that underlies so-called constraint-based algorithms for causal discovery such as PC/Fast Causal Inference (Spirtes et al., 2000) and Bayesian Constraint-based Causal Discovery (Claassen and Heskes, 2012a). Specialized variants, such as Cooper’s local causal discovery algorithm (LCD) (Cooper, 1997) and the Trigger algorithm (Chen et al., 2007), handle the case of three variables and are particularly relevant for our purposes. The statistical pattern in LCD takes a triplet of mutually dependent variables with the additional prior knowledge that one of the variables ($Z$) cannot be caused by the other two ($X$ and $Y$). As we will show in more detail in the Supplementary material, any causal model that now implies a conditional independence between the variables $Y$ and $Z$ conditioned upon $X$ has a causal link from $X$ to $Y$. So, reasoning backward, if we observe such a conditional independence in our observational data, we can interpret this as evidence for a causal link from $X$ to $Y$. A graph with a set of all possible causal models for three variables is presented in the Supplementary material. This causal pattern was first derived by Cooper in (Cooper, 1997), and later independently rediscovered in the context of genome biology in (Chen et al., 2007). This method has been applied in various papers in the biomedical research literature, such as (Karlson et al., 2007; Xia et al., 2010).
5.1.3 Related models and methodologies

LCD is closely related to other, arguably more standard approaches, such as Structural Equation Modeling, mediation analysis and instrumental variable analysis. Below we explain the similarities and differences between these methods.

Structural equation modeling

LCD, as most methods for causal discovery, is closely related to Structural Equation Modeling (SEM). Typically, SEMs are used in a confirmatory setting, where a limited amount of specific structures are taken into consideration and compared against each other by scoring them on the available data. Causal discovery methods are used in a more exploratory setting. They assume that there is some SEM underlying the data and then aim to reason about its structure. Under particular conditions, parts of the structure can be derived from conditional (in)dependencies (Spirtes, 2010). As explained in detail in the Appendix, LCD does exactly this for the specific case of three observed and possibly many latent variables. A key advantage of LCD over fitting different SEM structures to the data is that LCD automatically incorporates latent variables and then implicitly considers all possible models instead of just a few. This makes it possible for LCD to make generic statements about causal directions. PLS-SEM, for partial least squares structural equation modeling, is a specific variant of structural equation modeling (Monecke and Leisch, 2012; Hair Jr et al., 2016). Among others, it more explicitly handles latent variables and hence may be considered closer to the approach that we take in this chapter. However, also in PLS-SEM, one starts by specifying the structure between the (latent and measured) variables, which makes it different from LCD, which aims to infer the (invariant parts of the) structure from observational data. Thus, by using LCD we do not have to preselect several possible models to test, as typically done with PLS-SEM. As a result, LCD can potentially infer causal statements.

Mediation analysis

Mediation analysis starts from the assumption that the independent variable $Z$ (genetic factor) causes the dependent variable $Y$ (impulsivity/hyperactivity) and then aims to answer the question whether the effect of $Z$ on $Y$ can be (fully) explained by the mediator $X$ (inattention). The important difference with the analysis underlying LCD is that LCD does not start from the assumption that there is a causal relationship, but instead aims to derive one. Nevertheless, following the analysis detailed in the Supplementary material, it can be seen that...
we can only derive a causal statement if the data reveals a conditional independence, which amounts to one variable mediating the correlation between the other two.

**Instrumental variable approaches**

In so-called instrumental variable approaches (Angrist et al., 1996), the genetic factor $Z$ is called an instrument. It can be used to estimate the causal effect of the variable $X$ on the variable $Y$ in the presence of latent confounders. A valid instrument has to satisfy various criteria, among others that its effect on the variable $Y$ is fully mediated by the variable $X$ (in more complex settings possibly controlled for other variables). The main difference with LCD is that instrumental variable analysis starts from the assumption that there is a causal effect from $X$ to $Y$ and then tries to make use of the instrument $Z$ to estimate or bound its strength, whereas LCD uses the instrument $Z$ to try and infer the existence and direction of a cause-effect relationship between $X$ and $Y$, without attempting to estimate the causal strength of this relationship.

### 5.1.4 Goal

The goal of this study is to analyze whether such statistical patterns can be observed in studies of ADHD populations, and if so, what causal relationships these patterns then suggest. We will use symptom scores for inattention and hyperactivity/impulsivity as substitutes for the actual level of inattentiveness and hyperactivity/impulsivity. These then play the role of the variables $X$ and $Y$ above. For the variable $Z$ we will consider genetic variables such as gender and the DAT1 risk haplotype. These three variables clearly satisfy the premises of LCD: they are all mutually dependent (as shown in various other studies and easily checked for the data sets analyzed in this chapter) and it seems completely reasonable to assume that manipulations of inattentiveness and hyperactivity/impulsivity do not affect gender, nor the DAT1 risk haplotype.

### 5.2 Materials and Methods

#### 5.2.1 Materials

To infer causal relationships between ADHD symptoms we used three data sets, describing children, adolescents, and adults with ADHD. For each data set we only consider three variables: inattention symptom scores, hyperactivity/impulsivity symptom scores, and a genetic variable (either gender or a risk
5.2. Materials and Methods

haplotype). The main rationale for choosing these data sets is availability, as explained in more detail in the discussion.

The first data set was collected for the NeuroIMAGE project (von Rhein et al., 2014) (see www.neuroimage.nl) and considers adolescents. We will refer to this data set as the NeuroIMAGE data set. This data set includes N=903 participants (413 adolescents with ADHD, 228 unaffected siblings of ADHD probands, and 262 healthy control subjects) with a mean age of 16.7 years (min=5.7 years, max=28.6 years). The presence of ADHD symptoms was assessed by a semi-structured diagnostic interview Schedule for Affective Disorders and Schizophrenia for School-Age Children - Present and Lifetime Version (K-SADS-PL (Kaufman et al., 1997)) and Conners’ ADHD questionnaires from multiple informants (parents and children) (Conners et al., 1998). An algorithm was applied to create a combined symptom count from the interview and questionnaires (symptom range 0-18) (the algorithm is provided in (von Rhein et al., 2014)). Participants were diagnosed with ADHD if they met the full DSM-IV criteria for the disorder. For the current analyses, the sum of the symptom counts on the two symptom dimensions inattention (0-9) and hyperactivity/impulsivity (0-9) was used. In addition, we used the information on gender. In order not to complicate our analysis with ways to account for the dependencies between probands and their unaffected siblings, we ignore the siblings, leaving N=675 subjects in total. A more detailed description of the symptom assessment and recruitment process can be found in (von Rhein et al., 2014).

The second data set was collected by Peking University and is publicly available as part of the ADHD-200 Sample and parts of this data were described in several papers (Tian et al., 2008; Zhu et al., 2008; Cao et al., 2009; Wang et al., 2009b), () and considers children. We will refer to this data set as the ADHD-200 data set. This data set includes N=245 participants (102 children with ADHD, 143 control subjects) with a mean age of 11.7 years (min=8.1 years, max=17.3 years). The data set contains information about subjects’ ADHD symptom scores, disease status, gender, and IQ. Symptom scores were measured using the ADHD Rating Scale (ADHD-RS) IV (DuPaul et al., 1998), for which scores can range from 0 to 27 for each symptom domain. Also for this data set we will restrict our analysis to the two symptom scores and gender. We could not use the other data sets that are part of the ADHD-200 sample, because in those data sets the ADHD symptom scores were corrected for the effect of gender. More details about the ADHD-200 data sets are provided in (Cao et al., 2009).

The third data set was collected for the IMpACT project (Hoogman et al., 2013) and considers adults. We will refer to this data set as the IMpACT data set. This data set contains N=164 participants (87 adults with ADHD, 77 control subjects) with a mean age of 36.6 years (min=18.0 years, max=63.0
Chapter 5. Statistical evidence suggests that inattention drives hyperactivity/impulsivity in ADHD.

years. Subjects were assessed using the Diagnostic Interview for Adult ADHD (DIVA) (www.divacenter.eu). This interview focuses on the 18 DSM-IV symptoms of ADHD and uses concrete and realistic examples to thoroughly investigate whether the symptom is present now or was in childhood. In addition, a quantitative measure of clinical symptoms was obtained using the ADHD-DSM-IV Self Rating scale (Kooij et al., 2005), which has a range of scores from 0 to 9 for each symptom domain. To support the validity of the symptoms estimate based on self-reports, extra information about ADHD symptoms and impairment in childhood was obtained from parents and school reports, whenever possible. Patients were included in the study if they met the DSM-IV-TR criteria for ADHD in childhood as well as adulthood. As gender was not associated with ADHD in the adult data, we used an alternative genetic variable: the presence/absence of the DAT1 9/6 risk haplotype, a genetic polymorphism associated with ADHD in adulthood (Hoogman et al., 2013). More detailed information about the data collection and symptom assessment can be found in the original paper by Hoogman et al. (2013). For this type of analysis the use of DAT1 instead of gender as a genetic variable does not influence the validity of our results, since DAT1 also fulfills all the requirements of the LCD approach (DAT1 is correlated with inattention and hyperactivity/impulsivity, neither inattention nor hyperactivity can cause DAT1). This data set is described in more details in Chapter 4.

5.3 Data analysis

The inference of causal relationships from observational data crucially depends on the detectable absence and presence of conditional dependencies between variables (Pearl, 2000). For random variables that follow a multivariate Gaussian distribution, conditional independence corresponds to zero partial correlation. The partial correlation between \(X\) and \(Y\) given controlling variable \(Z\) is defined as the correlation between the residuals \(R_X\) and \(R_Y\) resulting from the linear regression of \(X\) with \(Z\) and of \(Y\) with \(Z\), respectively. In other words, partial correlation measures the degree of association between two random variables, with the effect of the controlling random variable removed. By measuring partial correlation it is possible to measure conditional independencies in the data.

Our symptom scores are not normally distributed and both gender and presence/absence of risk haplotype are binary variables. The standard approach of estimating conditional independencies uses Pearson partial correlation that relies on the assumption of Gaussian data. Since this assumption does not hold for our data, Pearson partial correlation is not guaranteed to represent conditional dependencies and independencies correctly for our data (Baba et al., 2004).
therefore replaced Pearson by Spearman rank partial correlation. Technically, a standard test for zero partial correlation with Spearman correlation instead of Pearson is valid for variables that obey a so-called non-paranormal distribution (Harris and Drton, 2013): a multivariate Gaussian distribution on latent variables, each of which is related to the observed variables through a monotonic transformation.

An alternative method to infer conditional independencies/dependencies from non-normally distributed data is to discretize the data at the risk of losing some statistical power and use the so-called Mantel-Haenszel test (Mantel and Haenszel, 1959). The basic idea of this test is to turn observed counts into expected counts under the assumption that there is a conditional independence and then check whether there is a significant difference between the expected and observed counts. For all three data sets we discretized the symptom scores into a binary variable using a median split, which had its threshold at 4.5. The observed counts were visualized in a cross table with a mosaic plot. A mosaic plot is an area-proportional hierarchical visualization of (typically observed) counts, composed of tiles (corresponding to the cells) created by recursive vertical and horizontal splits of a rectangle. The area of each tile is proportional to the corresponding cell entry given the dimensions of previous splits (Hartigan and Kleiner, 1981). Mosaic plots are excellent tools for visualizing conditional independencies: if two variables are conditionally independent given a third, this will show in the mosaic plot through straight lines as long as the conditioning variable is not represented at the lowest level of the hierarchy.

5.4 Results

We obtained consistent results for all three data sets. We provide a detailed description of the results for the ADHD adolescence data, including figures. A summary of the results for all three data sets is presented in Table 5.1.

In Figure 5.1 the NeuroIMAGE data set is displayed. It can be clearly seen that all three variables are significantly correlated (see Table 5.2 for correlations and effect sizes). Spearman’s partial correlation between gender and hyperactivity/impulsivity symptom scores conditioned upon inattention symptom scores is negligible (Spearman $R = -0.0008$, $p = 0.9826$). However, the Spearman partial correlation between gender and inattention symptom scores conditioned upon hyperactivity/impulsivity symptom is significantly different from zero (Spearman $R = 0.1235$, $p = 0.0013$). Spearman’s rank partial correlation coefficients are visualized in Figure 5.2.

The Mantel-Haenszel test for discretized data provided similar results. As shown in the mosaic plots in Figure 5.3, there is a significant difference ($\chi^2 = \ldots$
Table 5.1: Outcome of the conditional independence tests for the three different data sets. We check both whether inattention is conditionally independent of Gender/DAT1 given hyperactivity/impulsivity (second column) and whether hyperactivity/impulsivity is conditionally independent of Gender/DAT1 given inattention (third column). $R$ specifies the partial correlation (higher means more strongly correlated); chi-squared the Mantel–Haenszel test statistic (higher means larger deviation from independence). The p-values correspond to the null hypothesis that the two variables are conditionally independent.

<table>
<thead>
<tr>
<th>Type of test</th>
<th>Gender/DAT1 and inattention symptom scores conditioned upon hyperactivity/impulsivity</th>
<th>Gender/DAT1 and hyperactivity/impulsivity symptom scores conditioned upon inattention</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>NeuroIMAGE</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Partial correlation test</td>
<td>$R=0.1235$, $p=0.0013$</td>
<td>$R=-0.0008$, $p=0.9826$</td>
</tr>
<tr>
<td>Mantel-Haenszel test</td>
<td>Chi-squared $=11.37$, $p&lt;0.001$</td>
<td>Chi-squared $=0.15$, $p=0.70$</td>
</tr>
<tr>
<td><strong>ADHD-200</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Partial correlation test</td>
<td>$R=0.18$, $p=0.006$</td>
<td>$R=0.05$, $p=0.42$</td>
</tr>
<tr>
<td>Mantel-Haenszel test</td>
<td>Chi-squared $=10.98$, $p=0.001$</td>
<td>Chi-squared $=0.47$, $p=0.49$</td>
</tr>
<tr>
<td><strong>IMpACT</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Partial correlation test</td>
<td>$R=0.19$, $p=0.02$</td>
<td>$R=-0.01$, $p=0.91$</td>
</tr>
<tr>
<td>Mantel-Haenszel test</td>
<td>Chi-squared $=11.21$, $p=0.001$</td>
<td>Chi-squared $=0.005$, $p=0.95$</td>
</tr>
</tbody>
</table>

Figure 5.1: The NeuroIMAGE data set: Hyperactivity/impulsivity is plotted versus inattention symptoms for male and female. The bars indicate the histogram of the distribution. For visualization purposes random noise has been added to the discrete symptom scores.
Table 5.2: Correlation between the three variables for three data sets and the category of the effect size (in brackets). R represents Spearman rank correlation, and p-values correspond to the null hypothesis that the two variables are independent. Effect size estimates are based on the size of the correlation observed between two variables, where small, medium, and large correlation thresholds are respectively 0.10, 0.30, and 0.50 based on Cohen’s classification (Cohen, 1992).

<table>
<thead>
<tr>
<th></th>
<th>Gender/DAT1 and Inattention</th>
<th>Gender/DAT1 and Hyperactivity/Impulsivity</th>
<th>Inattention and Hyperactivity/Impulsivity</th>
</tr>
</thead>
<tbody>
<tr>
<td>NeuroIMAGE</td>
<td>R=0.187, p&lt;0.001 (small)</td>
<td>R=0.141, p&lt;0.001 (small)</td>
<td>R=0.759, p&lt;0.001 (large)</td>
</tr>
<tr>
<td>ADHD-200</td>
<td>R=0.288, p&lt;0.001 (small)</td>
<td>R=0.215, p=0.001 (small)</td>
<td>R=0.679, p&lt;0.001 (large)</td>
</tr>
<tr>
<td>IMpACT</td>
<td>R=0.307, p=0.001 (medium)</td>
<td>R=0.224, p=0.004 (small)</td>
<td>R=0.764, p&lt;0.001 (large)</td>
</tr>
</tbody>
</table>

Figure 5.2: Spearman’s partial correlation coefficients for the NeuroIMAGE data set representing inattention symptoms (In), hyperactivity/impulsivity (HI) symptoms, and gender (Gen). The bar colors represent the correlation value. Every cell(i,j) in the table shows Spearman partial correlation between two variables $X_i$ and $X_j$, conditioned on the remaining variables in the model. For example, figure shows that HI is independent of Gen given In (white square), while In depends on Gen given HI (pink square).
11.37, \( p < 0.001 \) between the observed and expected scores of inattention for the different genders, conditioned upon hyperactivity/impulsivity symptom level (Figure 5.3a). No significant difference (chi-squared=0.15, \( p=0.70 \)) is seen between the observed and expected scores of hyperactivity/impulsivity for different gender, conditioned upon inattention symptoms (Figure 5.3b). This implies that the triples in all three data sets satisfy the LCD-condition, i.e., where for a triplet of mutually dependent variables \( (X, Y, Z) \) with the prior knowledge that \( Z \) is not caused by \( X \) and \( Y \) we observe a conditional independency between \( Y \) and \( Z \) conditioned upon \( X \).

5.5 Discussion

The aim of this chapter was to apply a novel approach for causal discovery to improve our understanding of the strong correlation between the two symptom dimensions of ADHD. In three different and independent data sets, employing different instruments and raters to measure ADHD symptoms, and using different genetic variables, we found robust statistical evidence for a conditional independence of hyperactivity/impulsivity symptom level from a genetic variable, conditioned upon inattention symptom level. Without conditioning, the genetic variable (gender/risk haplotype) and hyperactivity/impulsivity were clearly dependent. Causal inference provides an explanation for this dependency cancellation: inattention causes hyperactivity/impulsivity.

5.5.1 Interpretation

The causal statement explaining the association between hyperactivity/impulsivity and inattention asks for a careful interpretation. Obviously, inattention as well as hyperactivity/impulsivity could be caused by many factors, directly or indirectly through yet other factors. What the causal model implies is that there is a significant causal path from inattention to hyperactivity/impulsivity, but not the other way around. Furthermore, there appears to be no (unobserved) factor with a similarly relevant causal path to both inattention and hyperactivity/impulsivity, since in that case the genetic variable and hyperactivity/impulsivity should be dependent conditioned upon inattention, which contradicts with the observed conditional independence. Summarizing the above, there are factors that influence inattention directly and influence indirectly hyperactivity/impulsivity via inattention. On the other hand there are also factors that influence hyperactivity/impulsivity directly, and have no effect on inattention. The variance of the hyperactivity/impulsivity explained by the inattention ranges between 67-77% for the data sets described in this study,
Figure 5.3: Mosaic plots of the observed counts for the NeuroIMAGE data set under the assumptions that a) hyperactivity/impulsivity symptom level and gender are conditionally independent given inattention symptom level; b) inattention symptom level and gender are conditionally independent given hyperactivity/impulsivity symptom level. The color of the cell represents the value of Pearson residuals of the Mantel-Haenszel test. Two variables are independent when the boxes proportions across categories are the same and there is a straight line that goes through these areas. For example, hyperactivity/impulsivity is independent of gender on Figure (a) when adjusted for the level of inattention, since there is no significant difference in the proportion of males and females for high and low level of hyperactivity. There is almost a straight line that divides high and low level of hyperactivity/impulsivity for both high and low level of inattention in Figure (a). Figure (b) shows that inattention depends on gender adjusted for the level of hyperactivity/impulsivity. There is significant difference in the proportion of females with high and low inattention when controlling for the hyperactivity/impulsivity level to the proportion of males with high and low inattention.
Chapter 5. Statistical evidence suggests that inattention drives hyperactivity/impulsivity in ADHD.

Figure 5.4: Causal relationships implied by our data for inattention (In), hyperactivity-impulsivity (HI), genetic variables (Gender or genotype), and behavioral estimates based on interview/questionnaire symptom scores.

based on the correlation between the two variables. The rest of the variance can be explained by factors that influence hyperactivity/impulsivity directly, not via inattention.

Note also that in this causal interpretation, we treat the outcome of the interviews/questionnaires as proxy for “inattention” and “hyperactivity/impulsivity”. In fact, “inattention” and “hyperactivity/impulsivity” themselves are perhaps best viewed as hidden concepts, which can be represented as latent variables that by themselves are linked to (causing) the respective symptoms. That we find this causal link between inattention symptoms and hyperactivity/impulsivity implies that there is likely to be a latent concept (which we may call “inattention”) that is quite accurately captured by the interview/questionnaire items related to inattention and which “causes” another latent concept (which we may call “hyperactivity/impulsivity”) that is quite accurately represented by items for hyperactivity/impulsivity in the interviews/questionnaires, see Figure 5.4. Furthermore, when we say that one variable “causes” another, we mean that if we manage to intervene on the first variable, this will change (the probability distribution of) the second variable. A similar subtle interpretation is implicit in many practical applications of causal discovery.
5.5.2 Related literature on ADHD

Early work on what we now know as ADHD in the 1940’s emphasized characteristics as hyperactivity and impulsivity as part of the so-called Minimal Brain Damage syndrome (Schwartz and Johnson, 1985). Later on, research failed to establish a firm link between hyperactivity and brain damage. Most children suffering brain damage did not develop hyperactivity, and fewer than 5% of hyperactive children appeared to suffer from brain damage (Rutter and Hersov, 1977). During the late 60’s and early 70’s, the focus shifted to problems in attention regulation. Virginia Douglas and her colleagues at McGill University in Canada were among the first to demonstrate the marked attention deficits seen in these children. Douglas argued that the major deficit was the inability to “stop, look, and listen” (Douglas 1972). After intense debate on what the primary features of the disorder were, the American Psychiatric Association published the DSM-III I 1980, and coined the disorder “Attention Deficit Disorder, with or without hyperactivity”. It was realized that the earlier diagnosis of hyperactivity in children does not necessarily mean that these children do not have inattention. It may well be that in small children, who have a more limited attention span than adults, inattention is harder to diagnose than hyperactivity. The results of Douglas’ research reflected the consensus that attention deficit, not hyperactivity, was the key to the disorder. The findings in our current analysis support this consensus.

The proposed model has many characteristics in common with the bi-factor model (Martel et al., 2010). The bi-factor model allows symptoms to be associated with general factors that are common for both symptoms, and specific factors for each symptom in particular. The model proposed in this chapter suggests that there are general factors that influence inattention and consequently hyperactivity, and specific factors that influence only hyperactivity. When given a causal interpretation, the bi-factor model explains a correlation between symptoms by a common cause (general factor), while our proposed model explains it by an effect from inattention to hyperactivity/impulsivity. Unfortunately, we cannot directly compare our study with the study in (Martel et al., 2010), which suggests that the bi-factor model outperforms other standard factor models of ADHD, since such analysis (Martel et al., 2010) requires symptom scores for each question, while in this study only aggregated scores per symptom were available. Furthermore, there are many slightly different variants that one could consider, each with various possible causal interpretations. In future work, we aim to extend the analysis of (Martel et al., 2010) on data with symptom scores for each question. Our current analysis strongly suggests to then explicitly incorporate gender or another genetic factor as an instrumental variable, since this
Statistical evidence suggests that inattention drives hyperactivity/impulsivity in ADHD.

Our causal model is in line with findings by Willcutt and coworkers (Willcutt et al., 2000) in a study of ADHD heritability in adolescent twin pairs. They showed that inattention is heritable for all levels of hyperactivity/impulsivity, whereas hyperactivity/impulsivity is heritable only when the level of inattention symptoms is high. This made the authors suggest that the etiology of hyperactivity/impulsivity is different in subjects who show a high level of inattention from that in subjects with low inattention. Such a hypothesis is perfectly consistent with our causal model: there are heritable factors that cause inattention and affect hyperactivity/impulsivity downstream of that, whereas those factors that lead to high hyperactivity/impulsivity do not necessarily lead to higher inattention. It has been found that hyperactivity/impulsivity symptoms remit more likely than inattention symptoms (Biederman et al., 2000). An obvious explanation, consistent with our model, is that those factors that directly affect hyperactivity become less prominent in adulthood, whereas the factors that affect hyperactivity through inattentions remain more or less constant. Longitudinal data are required to study such phenomena in more detail.

Considering clinical management of patients, the existence of a causal path from inattention to hyperactivity/impulsivity suggests that interventions (for example medication treatment) that decrease inattention are also likely to have a beneficial effect on the level of hyperactivity/impulsivity. On the other hand, interventions that affect hyperactivity/impulsivity cannot be expected to also have a positive effect on the level of inattention symptoms. This would further be consistent with reports that methylphenidate treatment of ADHD primarily targets attentional mechanisms by blocking the dopamine transporter in the striatum and the resulting increase in synaptic dopamine (Volkow et al., 2002).

5.5.3 Assumptions

As any statistical analysis, causal inference relies on several assumptions. Some of these assumptions are more fundamental, such as the assumption that we can use statistical tests to uncover the probabilistic (in)dependence relationships among the measured variables, and the assumption that reality can be properly modeled by acyclic Bayesian networks. These assumptions are discussed in detail in (Cooper, 1997). Note that we explicitly do not (have to) assume so-called causal sufficiency and hence do allow for the presence of latent confounders. These latent confounders could be clinical comorbidities or environmental mediators such as epigenetic mechanisms. Moreover, the fact that the observed conditional independencies were found in three independent data sets
representing three different age groups and considering two different control variables, appear to rule out that these results are an artifact of a selection bias.

The selection of the appropriate data sets for the analysis was based on previous findings in our research and the availability of the data. In Chapter 4 we describe a causal analysis of data from the IMpACT study on a larger number of variables. Here we noticed, among other things, the causal link between inattention and hyperactivity/impulsivity. The analysis in this chapter reveals that this causal link can also be found by restricting the analysis on the IMpACT data set to just three variables. To confirm this finding we considered the NeuroIMAGE and the ADHD-200 data sets. We did not have any other data sets available for the analysis that would satisfy the requirements mentioned in the introduction.

In this chapter the ADHD case-control sample was used instead of a random sample which raises the question whether a biased sampling plan will impact the empirical associations. To answer this question we checked how the results of the conditional independence tests change if we decrease the number of ADHD cases in the sample, keeping the number of controls the same. The tests showed that if the number of ADHD cases is very small (less than 10), the correlation between the gender and symptoms becomes insignificant, due to low variation in symptoms and small sample size. Consequently, a conditional independence test between inattention and gender, conditioned on hyperactivity also becomes insignificant. When we increase the number of ADHD cases the variation in symptoms in the sample increases as well as the sample size, making the correlation between gender and symptoms more pronounced. Consequently, the dependency between inattention and gender, conditioned on hyperactivity becomes significant. However, the dependency between hyperactivity and gender, conditioned on inattention does not depend on the number of ADHD cases and is always insignificant. This analysis implies that considering a random sample instead of an ADHD case-control sample, we obtain the same sets of conditional independencies provided that the sample size is large enough. We also repeated our analysis on the siblings from the NeuroImage data set, where we found evidence for the same pattern (not reported here, because statistically less significant than the other, larger data sets).

In this chapter we considered the division of ADHD into two symptom dimensions, namely inattention and hyperactivity/impulsivity. Other studies used a set of items that was much larger than the core ADHD symptoms and included items on mood, oppositional behavior and cognitive problems. These studies described a three-dimensional model splitting hyperactivity/impulsivity into separate dimensions of hyperactivity and impulsivity (Christiansen et al., 2011). Future studies may extend our current work into examining causal relationships between inattention, hyperactivity and impulsivity.
5.6 Conclusion

In this chapter we discuss the robust cancellation of dependency between hyperactivity/impulsivity and a genetic factor conditioned upon inattention observed in three different data sets. It is difficult to quantify one’s confidence in a statement such as “inattention causes hyperactivity/impulsivity”, if only because it strongly depends on the typical assumptions underlying causal inference. Some of these assumptions have been debated (see e.g., the discussion in (Glymour et al., 1999; Robins and Wasserman, 1999)), and some may claim that alternative approaches are more fruitful (e.g., causal inference as a missing-data problem as proposed by (Rubin, 1974); see however (Pearl, 2000)). It is clearly beyond the scope of this chapter to resolve such issues. We do argue that, when one is willing to apply these methods for causal inference (see e.g., (Karlson et al., 2007; Xia et al., 2010) for similar approaches within the biomedical domain), they suggest a logical explanation for the robust cancellation of dependencies in three different studies, which follows Ockham’s principle of parsimony to select the hypothesis with fewest assumptions. We further have discussed how such a causal model can be put in the historical context of the disease and may explain other findings such as those in (Willcutt et al., 2000) showing different etiology of the hyperactivity/impulsivity for subjects that have a high level of inattention from subjects with a low level of inattention. Last but not least, our causal model yields testable hypotheses, which may be validated in future intervention studies.

5.7 Appendix

5.7.1 Derivation of the LCD pattern

To explain the type of reasoning and underlying assumptions in more detail, we will spell out the LCD pattern (Cooper 1997) for the three variables that we are interested in: inattentiveness (‘In’), hyperactivity/impulsivity (‘HI’), and a genetic factor (‘Gen’), which can be either gender or a risk haplotype such as DAT1. We follow essentially the same reasoning as in (Cooper, 1997). An alternative proof can be found in (Chen et al., 2007).

Table 5.3 displays eight models, represented as so-called complete partial ancestral graph CPAGs (Zhang, 2008). This is an exhaustive representation of all possible models that fulfill conditions of LCD: 1. Inattention and hyperactivity/impulsivity are correlated with genetic factor; 2. Neither inattention, nor hyperactivity/impulsivity cause genetic factor. We only consider models that have at least two edges, since with less than two edges at least one of the vari-
ables will be independent of the other two, clearly violating the fact that all three variables are mutually dependent. Each CPAG, short for ‘complete partial ancestral graph’, by itself represents a whole class of possible causal models, not only over the observed variables but also over unknown latent variables. In these graphs, “X → Y” means that there must be a causal path from X to Y in the underlying causal model, “X ↔ Y” that there is no causal path from X to Y or from Y to X, so there must be a latent common cause affecting both X and Y. Circle marks are wild cards, that is, “X ◦→ Y” means either “X → Y” or “X ↔ Y”, and “X ◦← Y” any of “X → Y”, “X ← Y”, and “X ↔ Y” (note that here, in technical terms, we do allow for the possibility of latent variables, i.e., do not assume so-called causal sufficiency, but do assume that there is no selection bias and there cannot be any cycles; we will get back to these assumptions in the discussion). In our case, the assumption that no other observed variable in the model (neither inattention, nor hyperactivity/impulsivity) causes ‘Gen’ implies that when there is an edge between ‘Gen’ and, for example, ‘In’, it always comes with an arrowhead at ‘In’ and typically a circle mark at ‘Gen’. This implements our assumption that ‘In’ cannot cause ‘Gen’, without excluding the possibility that there is a latent common cause affecting the two (a circle mark at ‘Gen’). For example there can be a latent gene that causes both ‘In’ and ‘Gen’.

Reasoning forward, each CPAG now implies a set of (conditional) dependencies and independencies. They can be derived using a general property called m-separation (Richardson and Spirtes, 2002), here with just three variables arguably also through common sense. The implied (conditional) dependencies for each possible combination of variables are shown in the columns on the right, next to each of the CPAGs. Here, for example, “In ← HI | Gen” means that ‘In’ is independent of ‘HI’ when conditioned upon ‘Gen’. Graph (a) implies no (conditional) independencies at all. Graphs (b1) through (d2) are potentially more interesting: they all at least suggest one (conditional) independence. Letters refer to the models with the same set of edges, while numbers distinguish different edge directions in Table 5.3. For example, although model (b1) and (b2) represents two similar models with the same skeleton (an edge between ‘Gen’ and ‘HI’ and ‘Gen’ and ‘In’), and the same set of conditional independencies, they declare possibly different edge directions. Three causal models are possible a) causal effect from ‘Gen’ to ‘In’ and ‘HI’ (possible for both rows in Table 5.3 b1 and b2); b) causal effect from ‘Gen’ to ‘In’ and a common cause between ‘Gen’ and ‘HI’ (Table b1); c) causal effect from ‘Gen’ to ‘HI’ and a common cause between ‘Gen’ and ‘In’ (Table b2).

Now, if we observe a particular pattern of dependencies and independencies in the data, we can reason backward to tell which graph(s) can explain these. If indeed all variables are mutually dependent, graphs (b3), (c2), and (d2) drop out.
Chapter 5. Statistical evidence suggests that inattention drives hyperactivity/impulsivity in ADHD.

Table 5.3: Set of all possible models, represented as so-called CPAGs (3) that have at least two edges. Next to each graph a set of pairwise independencies and conditional independencies is represented. $X \perp \perp Y$ means that $X$ is independent of $Y$; $X \perp \perp Y \mid Z$ means that $X$ is independent of $Y$ conditioned on $Z$.

<table>
<thead>
<tr>
<th>CPAG</th>
<th>In $\perp$ HI</th>
<th>Gen $\perp$ HI</th>
<th>Gen $\perp$ In</th>
<th>In $\perp$ HI $\mid$ Gen</th>
<th>Gen $\perp$ HI $\mid$ In</th>
<th>Gen $\perp$ In $\mid$ HI</th>
</tr>
</thead>
<tbody>
<tr>
<td>(a)</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td>(b1)</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>Yes</td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td>(b2)</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>Yes</td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td>(b3)</td>
<td>Yes</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td>(c1)</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>Yes</td>
<td>No</td>
</tr>
<tr>
<td>(c2)</td>
<td>Yes</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>Yes</td>
<td>No</td>
</tr>
<tr>
<td>(d1)</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>Yes</td>
</tr>
<tr>
<td>(d2)</td>
<td>Yes</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>Yes</td>
</tr>
</tbody>
</table>
because they imply marginal independence between ‘In’ and ‘HI’. If all variables are mutually dependent, but we still have a conditional independence, graph (a) drops out, and only one of (b1), (b2), (c1), or (d1) applies, depending on which conditional independence holds true. If we find that ‘In’ and ‘HI’ are conditionally independent given ‘Gen’, we can conclude that ‘Gen’ must cause either ‘In’ or ‘HI’, but cannot tell which one. However, when one of the other two conditional independencies holds true, we have either (c1) or (d1), and in both cases we can infer a causal statement.

If according to the data ‘HI’ is independent of ‘Gen’ conditioned upon ‘In’ (model c1 in Table 5.3), LCD concludes that there must be a causal path from ‘In’ to ‘HI’ in any underlying causal model that can explain this particular pattern of (conditional) dependencies and independencies. This is exactly the pattern of conditional independencies/dependencies that was found in the three data sets discussed in the chapter.
A causal and mediation analysis of the comorbidity between ADHD and ASD

Autism spectrum disorder (ASD) and attention-deficit/hyperactivity disorder (ADHD) are often comorbid. In this chapter we explore the relationships between ASD and ADHD symptoms by applying causal modeling. We used a large phenotypic data set of 417 children with ASD and/or ADHD, 562 affected and unaffected siblings, and 414 controls, to infer a structural equation model using a causal discovery algorithm from Chapter 2. Three distinct pathways between ASD and ADHD were identified: (1) from impulsivity to difficulties with understanding social information, (2) from hyperactivity to stereotypic, repetitive behavior, (3) a pairwise pathway between inattention, difficulties with understanding social information, and verbal IQ. These findings may inform future studies on understanding the pathophysiological mechanisms behind the overlap between ASD and ADHD.

6.1 Introduction

Autism spectrum disorders (ASD) and attention-deficit/hyperactivity disorder (ADHD) are regarded as distinct disorders in the Diagnostic and Statistical Manual of Mental Disorders (DSM-5). ASD symptoms include impairments in interaction, communication and restricted, stereotyped, and repetitive behavior, whereas ADHD is characterized by symptoms of inattention and hyperactiv-

This chapter is based on Sokolova et al. (2017), “A causal and mediation analysis of the comorbidity between attention deficit hyperactivity disorder (ADHD) and autism spectrum disorder (ASD)”, published in the Journal of Autism and Developmental Disorders.
ity/imulsivity (Association, 2013). In previous versions of the DSM, ASD was an exclusion criterion to be diagnosed as having ADHD. As a result, these disorders were studied separately from each other for many years. However, recent research recognizes considerable clinical, genetic, and neuropsychological overlap between ASD and ADHD (Rommelse et al., 2011, 2010) and within the DSM-5, ADHD can now be diagnosed in conjunction with ASD. Various studies showed that 22–83% of children with ASD have symptoms that satisfy the DSM-IV criteria for ADHD (Ronald et al., 2008; Matson et al., 2013), and vice versa, 30–65% of children with ADHD have clinically significant symptoms of ASD (Clark et al., 1999; Ronald et al., 2008). In clinical practice, it is sometimes difficult to differentiate between ASD and ADHD, partly due to the entanglement of symptom descriptions of both disorders (Luteijn et al., 2000). This might explain why a substantial proportion of children have been alternatively given a diagnosis of one or the other disorder throughout development (Fein et al., 2005).

A strong body of twin-, family-, and linkage studies have consistently shown that ASD and ADHD share a portion of their heritable etiology (Lichtenstein et al., 2010). About 50-72% of the contributing genetic factors overlap between ASD and ADHD (Lichtenstein et al., 2010; Rommelse et al., 2010). Furthermore, similar deficits in executive function, social cognition, and motor speed have been linked to both ASD and ADHD (see for an extensive review, (Rommelse et al., 2011)). Relationships between ASD and ADHD appear to be stronger during certain developmental periods than others, with rather strong ASD/ADHD constellations during adolescence and weaker correlations in early childhood and at adult age. This might be due to that optimal social adaptation and EF skills matter most in adolescence (Hartman et al., 2016).

The main goal of this chapter is to investigate what is now needed to resolve this issue of symptom entanglement and alternating diagnoses. Some studies have tried to examine to which degree different symptom domains cluster together, and to which extent these domains are caused by the same genetic and environmental influence (Polderman et al., 2013; Ronald et al., 2014; Taylor et al., 2015). It has been proposed that the association between ASD and ADHD traits is primarily due to shared attention-related problems (inattention and attentional switching capacity), suggesting that biological pathways involving attentional control may be a key factor in unraveling the genetic causes of these disorders (Polderman et al., 2013). However, it is controversial to assume that attentional switching deficits belong solely to ASD and not ADHD. Impulsivity and inattention are often present in individuals with symptoms of ASD and these symptoms have a strong phenotypic and genetic overlap with non-social autistic traits, such as repetitive behavior (Ronald et al., 2014). In contrast, another study showed that genetic overlap was strongest between communication
difficulties typical of ASD and ADHD, while repetitive behavior and social difficulties showed only moderate genetic overlap (Taylor et al., 2015). Thus, these studies provide different explanations of comorbidity between ADHD and ASD.

These studies did not assess whether or not the observed links between specific ASD and ADHD traits were due to direct associations or indirect associations. That is, whether or not traits are correlated due to the causal effect of one variable on another or an unobserved common cause for both traits (direct paths) or due to an indirect association mediated via another trait (indirect paths). For example, the finding that social problems were only moderately correlated with hyperactivity, yet strongly correlated with inattention (Ronald et al., 2014), may suggest that the former correlation is explained by an indirect path from social problems to hyperactivity mediated via inattention. Being able to differentiate between direct and indirect paths may greatly improve our understanding of the co-occurrence of ASD and ADHD. In clinical practice it is often unclear what amplifies what, i.e., whether the ADHD related impulsivity is causing the social problems, or reversely, whether the repetitive behaviors are mistaken for hyperactivity. Answering these questions of direction and causation may have significant clinical implications, as it may inform therapeutic interventions.

Standard research methods such as correlation analysis or clustering do not provide the possibility to infer directionality from cross-sectional data. In the current study, the aim is to build a causal model describing the direction of the associations between specific behavioral symptoms of ASD, ADHD, and general factors via a structural equation model (SEM), using the Bayesian Constraint-based Causal Discovery (BCCD) algorithm (Claassen and Heskes, 2012a). This is an exploratory approach that learns the structure of a SEM from the observed data instead of the more commonly published confirmatory approach that tests a priori defined hypothetical networks. The idea of exploratory structure learning algorithms (Pearl, 2000) is based on the connection between conditional independencies and causal relationships. Thus, by finding conditional independencies in cross-sectional data, it is possible in particular cases to infer parts of the structure of a SEM and make (preliminary) predictions about causation. BCCD infers the skeleton of the SEM that describes direct associations as well as the direction of effects from data (a detailed description is provided in the Appendix). While the skeleton can be accurately inferred from a relatively small sample size, the accurate inference of causal directions requires larger sample sizes (Claassen and Heskes, 2012a) and the presence of particular patterns to be able to infer the directions. As a second step, standard mediation analysis is applied to test direct or indirect relationships obtained through causal modeling.

In sum, our aim is to explore the relationships between specific ASD and ADHD symptoms by applying causal modeling to a large set of observed data.
(n=1393) including children with ADHD and/or ASD, their siblings and control children. Some generic factors are included in our analysis that are known to be associated with ASD and ADHD, namely age, gender, and IQ (Gardener et al., 2009; Mill and Petronis, 2008). The current approach primarily determines whether the association between variables is direct, rather than determining the direction of this association, but inferred directions are also included as preliminary hypotheses that should be further tested in independent samples.

6.2 Materials and Methods

6.2.1 Participants

Participants from two large-scale family-genetic studies, the Biological Origins of Autism (BOA, data collected between 2008-2012) study and the Dutch part of the International Multicenter ADHD Genetics (IMAGE data collected between 2004-2008) study (van Steijn et al., 2012), were included in the current study. Inclusion criteria for all participants were at least two biological siblings (in case of families: at least one child with a clinical diagnosis of ASD or ADHD), offspring age between 4 and 20 years, European Caucasian descent, offspring IQ \( \geq 70 \), and no diagnosis of epilepsy, brain disorders, or known genetic disorders, such as Down-syndrome or Fragile-X-syndrome.

All participants were carefully phenotyped for ASD and ADHD using validated and standardized questionnaires and diagnostic interviews. Briefly, both the children already clinically diagnosed with ASD and/or ADHD, their siblings, and the control children were screened for the presence of ASD and ADHD symptoms using the parent-reported Social Communication Questionnaire (SCQ) (Rutter et al., 2003) and the parent- and teacher–reported Conners Rating Scales-Revised (CPRS; CTRS), respectively (Conners, 1996). Raw scores of \( \geq 10 \) on the parent-rated SCQ Total score, \( \geq 15 \) on the teacher-rated SCQ Total score, and T-scores \( \geq 63 \) on the Conners’ DSM-IV Inattention, Hyperactivity-Impulsivity, or Combined scales were considered as clinical. A lower cutoff for the parent reported SCQ to avoid false negatives in their undiagnosed offspring (Corsello et al., 2007). All children scoring above cut-off on any of the screening questionnaires underwent full clinical ASD and ADHD assessment, including the Autism Diagnostic Interview-Revised (ADI-R) structured interview for ASD (Le Couteur et al., 2003) and the Parental Account of Childhood Symptoms ADHD subversion (PACS) for ADHD (Taylor, 1986). Control children were required to obtain non-clinical scores (i.e., a raw score < 10 on the SCQ and T-score < 63 on both parent and teacher reported CRS-R DSM-IV scales) in order to be accepted in this study.
The total sample contained 1393 participants, including 586 patients (317 ADHD only, 130 ASD only, and 139 combined ASD+ADHD), 393 unaffected siblings, and 414 controls. Demographics of the study sample are shown in Table 6.1. A more detailed description of participant selection can be found in (van Steijn et al., 2012; Oerlemans et al., 2014).

6.2.2 Measures

To apply causal discovery using the BCCD algorithm, the following variables were selected.

- Age of the participant
- Gender
- Current ADHD symptoms assessed with the parent and teacher reported CRS-R scales
- Inattention symptoms (CRS DSM-IV inattention subscale)
- Hyperactivity symptoms (hyperactivity items of the CRS DSM-IV hyperactivity/impulsivity subscale)
- Impulsivity symptoms (impulsivity items of the CRS DSM-IV impulsivity subscale)
- Current ASD symptoms assessed with four subscales of the parent-reported Child Social Behavior Questionnaire (CSBQ) (Hartman et al., 2015). A full list of CSBQ items is provided in Appendix. For clarity we provide a few examples items for each symptom type of CSBQ.
- Reduced contact and social interests (Has little or no need for contact with others, makes little eye contact, etc.)
- Difficulties in understanding social information, referred to as social inceptness further in the text (Takes things literally, e.g., does not understand certain expressions, Does not fully understand what is being said, i.e., tends to miss the point, etc.)
- Fear of/and resistance to changes (Remains clammed up in new situations or if change occurs, panics in new situations or if change occurs, etc.)
- Stereotyped, repetitive behavior (Constantly feels objects, smells objects, etc.)
# Chapter 6: A Causal and Mediation Analysis of the Comorbidity Between ADHD and ASD

Table 6.1: Demographics of the study sample. The standard deviation of the average value is indicated in brackets.

<table>
<thead>
<tr>
<th>Row Labels</th>
<th>Controls</th>
<th>Siblings</th>
<th>Patients</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Total N</strong></td>
<td>414</td>
<td>116</td>
<td>24</td>
</tr>
<tr>
<td><strong>% Male</strong></td>
<td>42%</td>
<td>59%</td>
<td>75%</td>
</tr>
<tr>
<td><strong>Age in years</strong></td>
<td>11.20 (min= 4.33, max= 20.08)</td>
<td>10.84 (min=5.47,max=20.2)</td>
<td>11.49 (min=7.34,max=17.75)</td>
</tr>
<tr>
<td><strong>Inattentiveness raw score (minimum level=0, maximum level=27)</strong></td>
<td>2.54 (2.67)</td>
<td>13.83 (4.72)</td>
<td>14.90 (4.91)</td>
</tr>
<tr>
<td><strong>Hyperactivity raw score (minimum level=0, maximum level=18)</strong></td>
<td>1.13 (1.32)</td>
<td>6.71 (3.38)</td>
<td>7.27 (2.51)</td>
</tr>
<tr>
<td><strong>Impulsivity raw score (minimum level=0, maximum level=9)</strong></td>
<td>0.63 (0.88)</td>
<td>3.46 (1.92)</td>
<td>4.42 (2.02)</td>
</tr>
<tr>
<td><strong>Reduced contact and social interests (minimum level=0, maximum level=22)</strong></td>
<td>0.94 (1.72)</td>
<td>3.52 (3.77)</td>
<td>10.62 (4.81)</td>
</tr>
<tr>
<td><strong>Social ineptness (minimum level=0, maximum level=14)</strong></td>
<td>1.41 (1.74)</td>
<td>5.63 (3.54)</td>
<td>8.95 (3.57)</td>
</tr>
<tr>
<td><strong>Fear of and resistance to changes (minimum level=0, maximum level=6)</strong></td>
<td>0.33 (1.04)</td>
<td>1.49 (3.36)</td>
<td>3.24 (3.07)</td>
</tr>
<tr>
<td><strong>Repetitive behavior (minimum level=0, maximum level=15)</strong></td>
<td>0.45 (0.84)</td>
<td>2.72 (1.68)</td>
<td>5.05 (1.89)</td>
</tr>
<tr>
<td><strong>Verbal IQ</strong></td>
<td>107.64 (13.16)</td>
<td>99.39 (13.82)</td>
<td>94.43 (15.31)</td>
</tr>
<tr>
<td><strong>Performance IQ</strong></td>
<td>105.82 (14.25)</td>
<td>102.59 (13.91)</td>
<td>101.98 (18.53)</td>
</tr>
</tbody>
</table>

Note: ADHD = Attention Deficit Hyperactivity Disorder, ASD = Autism Spectrum Disorder.
6.2. Materials and Methods

- Intelligence as measured using the Wechsler Intelligence Scale for Children (WISC-III) or the Wechsler Adult Intelligence Scale (WAIS-III), depending on child’s age (Wechsler, 2002, 2000)
- Verbal IQ, prorated by subtests Similarities and Vocabulary
- Performance IQ, prorated by the subtests Block Design and Picture Completion

In this study we considered the raw data ADHD symptoms for our analyses instead of the T-scale score, since T-scale scores are adjusted for the effect of gender and age. BCCD can model the effect of age and gender into account, and so we avoided unwanted ‘double correction’. Moreover, we separated impulsivity and hyperactivity subscales based on item scores, instead of using the ‘standard’ DSM hyperactivity/impulsivity subscale to examine the effect of each specific trait. For ADHD symptoms, scores assessed by parents and teachers were provided. To increase the reliability of the symptom assessment, for each subject we averaged the symptom scores from parent and teacher. The main two reasons for that are: (1) ADHD is diagnosed when several symptoms are prevalent in at least two or more settings, thus many clinicians find that parent and teacher ratings are helpful in the diagnostic process; (2) parent- and teacher ratings are highly correlated (R=0.64, p<0.0001), which makes it difficult to compare them independently. Unfortunately, it was not possible to obtain information about ASD symptoms from a second observer, thus ASD symptoms were assessed only based on parents report.

The CSBQ contains items refer directly to the DSM-IV criteria for autistic disorder, but also represent less severe variations of these criteria as well as ASD-associated problem such as executive function problems and disruptive behavior in social settings (Hartman et al., 2012). We opted for the CSBQ instead of the SCQ to assess ASD symptoms, because we were specifically interested in current behavior, whereas the SCQ mainly refers to behavior at age 4-5 years. Multiple studies have shown that the CSBQ has good psychometric properties with regard to test-retest and interrater reliability, internal consistence of the scales (all reliability indices > .75), and good criterion validity both for high-functioning children and for children with mild to moderate mental retardation (Hartman et al., 2006; de Bildt et al., 2009; Noordhof et al., 2015; Jaspers et al., 2013; Greaves-Lord et al., 2013). For ASD symptoms only parent scores were provided, so it was not possible to combine them with teacher scores. The reason that only parent-reported ASD symptoms were included in the study is that there is no teacher version of the CSBQ available. The selected WISC-
III/WAIS-III subtests are known to correlate between 0.90 and 0.95 with the full-scale IQ (Groth-Marnat, 1997).

6.2.3 Data analyses

In the first step of the analysis a causal discovery algorithm was used to learn the structure of a SEM, and to formulate hypotheses about direct and indirect relations between variables. Appendix describes the link between SEM and causal discovery, provides a description of existing algorithms for causal discovery as well as our motivation for using Bayesian Constraint-based Causal Discovery (BCCD) (Claassen and Heskes, 2012a). A detailed description of BCCD is also provided in Chapter 2. This algorithm infers statements representing causal relationships and estimates the reliability of these statements. The method outputs information about potential interactions between observed variables and does so in two ways: through the skeleton and through orientation. The skeleton describes mediation: two variables are connected if the association between them is not mediated by any other observed variable. Tails and arrows provide information about the direction of the association. The results are visualized through a causal graph by considering statements with reliability higher than 50%. Our method can incorporate prior knowledge about the domain. Here the assumption that gender and participant’s age cannot be caused by any other observed variables in the model was implemented, since chronologically the former are present in the lifespan earlier than the latter.

In the second step of the analysis standard mediation analysis was applied to explicitly check some of the hypotheses generated by our causal analysis. Mediation analysis distinguishes between independent variable, dependent variable, and potential mediators (Baron and Kenny, 1986). To test whether the effect of the independent variable is indirect, a regression model was built that aimed to predict the dependent variable from the independent variable and potential mediators. If the regression coefficient was statistically significant for the potential mediators, but not for the independent variable, a conclusion can be made that there was not enough evidence to reject the hypothesis that the effect of the independent variable is indirect.

Note here that the same data is used twice: to generate hypotheses and to test them. Consequently, the reported p-values of the mediation test should be treated with care. These p-values only indicated the significance as if the specific hypothesis had been coined prior to observing any data. Also note that the data set contains siblings from the same family. To test whether there is an effect of familiality on the resulting causal model, a sensitivity analysis using a subsample of singletons was performed, including only one subject per family.
6.3 Results

Due to the reduction of the sample size the reliability of the causal links tends to drop. However, if the model is stable, the main links will be preserved in the new model.

Running the BCCD algorithm we inferred reliability estimates of the causal relations between variables (Appendix Tables 6.2, 6.3, 6.4) and built a graph summarizing these relationships presented in Figure 6.1. In this graph an edge between two variables suggests that no other variable in the model can make these variables independent, which we call here a direct relationship. This can be either an effect of one variable on another (“$A \rightarrow B$”), unobserved common cause (“$A \leftrightarrow B$”) or a selection bias (“$A - B$”). If the direction of an edge between two variables is uncertain it has a circle mark “○”.

The general structure of the network matches other studies in the literature: gender influences symptom counts with males having higher scores than females (Cantwell, 1996; Ramtekkar et al., 2010); age influences hyperactivity level with older subjects having lower level of hyperactivity than younger subjects (Biederman et al., 2000); ADHD symptoms are associated with ASD symptoms (Ronald et al., 2008) and both are associated with IQ (with children having ASD, ADHD, or both having lower IQs in general than children without the disorder) (Vaida et al., 2013). Moreover, both ASD and ADHD symptoms are strongly interconnected, resulting in a separate cluster (also called a clique: a complete subgraph, in which all variables are pairwise interconnected) of ADHD and ASD symptoms. The same holds for IQ.

The inferred network suggests that the ASD traits ‘social ineptness’ and ‘stereotyped, repetitive behaviors’ are directly and differentially associated with ADHD symptoms. Social ineptness is associated with inattention and impulsivity, while stereotyped, repetitive behavior is associated with hyperactivity (but not impulsivity). Our network also shows that verbal IQ is a linking factor between ADHD and ASD, since there is a link from verbal IQ to both ADHD and ASD symptom traits. To get a better understanding of these observed direct associations, we zoom in on each link. The direction of these causal links contains circle marks, indicating uncertainty in the causal directions. For example, a link ‘○→’ between impulsivity and social ineptness and between hyperactivity and repetitive behavior is either a causal link or a selection bias. A link ‘○→’ between inattention and social ineptness is either a causal link or an unobserved common cause.

Based on the inferred model, there is a direct association of the social ineptness with inattention and impulsivity. Both links have a very strong reliability
Chapter 6. A causal and mediation analysis of the comorbidity between ADHD and ASD

Figure 6.1: Causal model representing causal relationships between variables in our combined ADHD and ASD data set. Edge directions represent either a causal effect (“A → B”), an unobserved common cause “A ↔ B” or a selection bias “A − B”. Non-identifiable edge directions are marked with a circle mark “○”. Notation “A − ○ B” is either a causal effect “A → B”, or a selection bias “A − B”; “A○ → B” is either a causal effect “A → B”, or a common cause “A ↔ B”; “A ○ − ○ B” is either a causal effect “A → B” or “A ← B”, a selection bias “A − B” or a common cause “A ↔ B”. No edge between variables means that these variables are conditionally independent given the other variables in the network. Reliability estimates for the presence of an edge are depicted as percentages. Direct links between ASD, ADHD, and IQ are marked in red.
6.3. Results

Figure 6.2: Regression model for mediation analysis that predicts dependent variable (in grey) social ineptness using inattention and impulsivity as a mediator and hyperactivity as independent predictor. The regression model is presented at the top of the figure, the significance of the regression coefficient is shown next to the edge.

for a direct link (>99%), providing strong evidence of a direct association. Mediation analysis confirmed that there was no direct link between hyperactivity and social ineptness ($\beta = 0.01, p = 0.79$) (Figure 6.2). We provide the first figure of the regression analysis as an example in the main text, other figures of this type of analysis can be found in the Appendix.

Another direct association between ASD and ADHD traits can be seen between hyperactivity and stereotyped, repetitive behaviors (reliability for direct link >99%). No direct causal links are found between repetitive behavior and inattention or impulsivity (Figure 6.1). Mediation analysis confirmed that there are no direct paths between inattention and repetitive behavior ($\beta = 0.03, p = 0.11$), and impulsivity and repetitive behavior ($\beta = 0.01, p = 0.86$), but an indirect one mediated through social ineptness, which may explain the correlations observed between these variables (Appendix Figure 6.3).

Our analyses also indicated that inattention and social ineptness are associated via verbal IQ due to direct links between inattention and IQ (reliability for direct link >79%), and between social ineptness and IQ (reliability for direct link >99%). Taking into account the link between inattention and social ineptness mentioned before, all three variables are pairwise connected, which can be a sign of an unobserved common cause for these variables. Mediation analysis showed that verbal IQ is only indirectly associated with impulsivity
Chapter 6. A causal and mediation analysis of the comorbidity between ADHD and ASD

(\(\beta = 0.09, p = 0.74\)) and hyperactivity (\(\beta = -0.25, p = 0.16\)) and that this is mediated through inattention (Appendix Figure 6.4). Mediation analysis also revealed that there is no direct link between verbal IQ and repetitive behavior (\(\beta = -0.27, p = 0.12\), reduced contact (\(\beta = 0.02, p = 0.87\), and fear of change (\(\beta = 0.04, p = 0.89\), but that these effects are mediated through social ineptness and hyperactivity (Appendix Figure 6.5).

Our model makes preliminary predictions about the directions of the causal links between ASD and ADHD traits. According to the model, hyperactivity may have a causative effect on repetitive behavior, with reliability of the link direction >91%. The direction of this link is inferred from 1) the assumption that hyperactivity does not cause age, and 2) the dependency between age and stereotypic behavior became insignificant when controlling for hyperactivity (\(R = -0.01, p = 0.86\)). Moreover, our model indicates that impulsivity may have a causative effect on social ineptness (and not visa-versa) with reliability >85%. This direction is inferred from 1) the assumption that impulsivity does not cause age, and 2) the dependency between age and social ineptness became insignificant when controlling for impulsivity (\(R = 0.01, p = 0.79\)).

6.4 Discussion

In the current study we applied exploratory causal modeling to investigate the co-occurrence of ADHD and ASD by incorporating their core symptom domains into a single integral model. Since ASD and ADHD symptom domains are all significantly pairwise correlated, raw correlation-based methods would not provide any insight into the direct and indirect association between these symptom domains. The causal method applied in this chapter builds a more complete model, distinguishes between direct and indirect associations, and allows us to make preliminary predictions about causation. These predictions were corroborated by mediation analysis. The results suggest at least three separate pathways between ADHD and ASD: a) a pathway from impulsivity to social ineptness, and b) a pathway from hyperactivity to stereotyped behavior c) a cluster of inattention, social ineptness and verbal IQ, with a possible common cause.

Our findings suggest that there are multiple distinct pathways and causes for the co-occurrence between ASD and ADHD. The strongest link was found between social communication difficulties, inattention and impulsivity. This corroborates previous reports based on both cross-sectional (Polderman et al., 2014) and longitudinal data (St Pourcain et al., 2014) that part of the association between ASD and ADHD may be due to shared attention-related problems. This is also in accordance with the outcome of reviews by our group (Visser et al., 2016) as well as others (Jones et al., 2014) that attentional problems at a very
early age may precede the onset of clinical manifestations of ASD, ADHD, or both disorders. These attentional problems may include, for example, problems in attentional shifting and disengaging impairments (Jones et al., 2014; Visser et al., 2016). As a novel finding, our model putatively suggests that impulsivity has a causative effect on social ineptness. Such a causal link from impulsivity to social ineptness would make intuitive sense. To interact effectively with others, an individual must be able to control impulsive behaviors. Impulsive symptoms may lead a person to miss social cues, for example, because they act prematurely or interrupt the other person (Leitner, 2014), which in turn may result in social difficulties. The relevance of impulsivity is reflected in cognitive studies that describe deficits in executive functioning in young children with ADHD and/or ASD, as measured in tests of response inhibition and interference control.

Our model does not make (preliminary) predictions on the causal direction between inattention and social ineptness. It does put inattention and social ineptness in one cluster with verbal IQ, which can be an indication of an unobserved common cause between these variables, for example a shared genetic factor. Verbal IQ refers to the capacity to use language in order to express oneself, comprehend stories, and understand other people, but also to self-directed speech that supports self-control. Previous studies have reported on language problems in both ASD and ADHD (Geurts and Embrechts, 2008; Geurts et al., 2004; Leonard et al., 2011). Children with ASD often have a delayed development of spoken language, fail in normal back-and forth conversations, and use language in a stereotypic and repetitive manner. The diagnostic criteria for ADHD also include behaviors suggesting social-communication dysfunction, such as talking excessively, interrupting others, and not listening to what is being said (Association, 2013). These communication deficiencies may contribute to social interaction problems that are typical for individuals with ASD and ADHD. A number of studies have reported on chromosomal regions that may harbor quantitative trait loci (QTLs) for language and communication problems in ASD, including chromosome 7q (Alarcon et al., 2002), which was also identified in a study looking for potential pleiotropic loci for ASD and ADHD (Nijmeijer et al., 2010). Nijmeijer et al. (2010) also found suggestive linkage on chromosome 15q for the SCQ communication subscale in their sample of ADHD families. Furthermore, relatively poor verbal comprehension is more often found in children with ASD (Charman et al., 2011; Rundblad and Annaz, 2010). Further study is needed to increase our knowledge on possible pleiotropic (genetic) risk factors that underlie the complex associations between inattention, social ineptness, and verbal competence.

A second pathway identified was between hyperactivity and repetitive behavior. In most studies, impulsivity and hyperactivity are regarded as one com-
bined feature, but our results suggest that these symptoms may be differentially associated with ASD symptoms. Some studies have previously reported on the link between repetitive behaviors and hyperactivity (Polderman et al., 2014; Gabriels et al., 2005; Ronald et al., 2014; Polderman et al., 2013). It has been argued that repetitive behavior and ADHD are due to a lack of inhibitory control, but contrasting findings have also been reported (Rommelse et al., 2011). Our model putatively suggests that individuals who are hyperactive and therefore less able to inhibit motor behaviors may, as a result, engage also more often in various motor behaviors that are classified as stereotypic, such as flapping arms/hand when excited or making odd and fast movements with fingers or hands (all items from the CSBQ ‘stereotypic behavior’ subscale). However, Polderman et al. (2014) proposed that the association may be conversely explained by repetitive behaviors interfering with the ability to switch attention from one task to another. Furthermore, inhibitory control is also associated with impulsivity, which was only indirectly related to repetitive behaviors according to our model. Further research on the direction of the link between hyperactivity and repetitive behavior is therefore needed.

Our putative predictions about the causal directions in the two pathways between ADHD and ASD (from ADHD inattention/impulsivity to ASD social ineptness, and from ADHD hyperactivity to ASD stereotyped, repetitive behavior) suggest that interventions that decrease inattention/impulsivity related difficulties are also likely to have a beneficial effect on social functioning, but not the other way around; interventions that affect social functioning cannot be expected to also have a positive effect on the level of inattention/impulsivity. The same logic is applicable for the effect of hyperactivity on repetitive behavior. These findings are consistent with results from longitudinal study by St Pourcain et al. (2011). They showed that children with high probability for persistent hyperactive-inattentive symptoms had a high probability for persistent social communication deficits, but not vice versa (St Pourcain et al., 2011). Our results may also fit well with the gradient overarching disorder theory, which proposes that ADHD is a less severe subtype within the ASD spectrum (van der Meer et al., 2012). As a consequence, individuals with (more severe forms of) ADHD are also highly likely to have increased (sub)clinical levels of ASD symptoms. It is important to note however, that our findings are based on just one (albeit rather large) sample, which needs to be replicated in other, independent samples, ideally in a longitudinal design.

Several strengths and limitations should be taken into account when assessing the results of the current study. The main strength of this study is the application of a novel causal discovery method for data analysis. This method considers all variables together, infers both direct and indirect dependencies between
variables, provides a reliability measure for each edge in the network, and is able to detect latent common causes. This method does not require longitudinal or interventional data and can infer causal statements based on cross-sectional data (Pearl, 2000). Another strength is the use of a large, carefully phenotyped sample of affected and unaffected siblings and control children, allowing us to study the full spectrum of ASD and ADHD symptoms.

A limitation of our causal discovery method is that it is an exploratory analysis – it provides new hypotheses that need to be tested using other methods and requires independent replication through experiments or additional data. Another limitation of our study is that the conclusions mainly apply to individuals with average IQ, as we excluded participants with an IQ below 70. This is not representative of the ASD population at large that includes a considerable proportion of individuals with ASD with an intellectual disability. Furthermore, we excluded individuals with known epilepsy, brain disorders, or genetic syndromes, and who were not of European Caucasian descent. Thus, caution is warranted when interpreting our results. In addition, including data of genetically related individuals may cause interpretation problems, due to the possibility of unobserved latent associations between variables that were not taken into account. We tackled this problem by running a sensitivity analysis using a subsample of singletons - including only one subject per family - to evaluate the impact of familiarity. We obtained a highly similar network structure with only a few missing edges due to reduced statistical power as a consequence of the reduction of the sample size by half, indicating the robustness of our approach and our findings (Figure 6.6 in Appendix).

In conclusion, our results indicate that the often reported co-occurrence of ASD and ADHD might be explained by three distinct pathways: a) between inattention/impulsivity and social ineptness, and b) between hyperactivity and stereotypic, repetitive behaviors c) through verbal IQ. These findings may inform future studies on understanding the (pathophysiological) mechanisms behind the overlap between ASD and ADHD.

6.5 Appendix

6.5.1 Causal modeling

One of the most popular and intuitive ways to represent causal models in the social sciences is through structural equation modeling (SEM) (Beran and Violato, 2010). One way of working with SEMs is to provide several hypothetical networks based on some prior knowledge and then compare these networks based on a particular metric (for example the AIC or BIC score). This is a confirmatory
analysis that is used to test a particular hypothesis, which is often applied in, for example, twin studies. Another approach is to try and learn the structure of SEM from the observed data. The basic idea of structure learning algorithms is described in Turing award winner Judea Pearl’s work (Pearl, 2000) that shows a connection between conditional independencies and causal relationships. Thus, by learning conditional independencies in cross-section data, it is possible in particular cases to learn the structure of a SEM and to make predictions about causation. Causal modeling is primarily an exploratory approach aiming to find novel causal paths that were not known in advance. In this chapter we consider the second approach, since our goal is to explore links between ADHD and ASD traits rather than to confirm known links.

The two main approaches to learn the structure of a SEM from data are the so-called score-based and constraint-based approaches (Daly et al., 2011; Pearl, 2000). The score-based approach provides a measure of reliability of the inferred causal network which makes the interpretation of the results easier and prevents incorrect categorical decisions (Heckerman et al., 1999). However, this approach often relies on the assumption that there are no common confounders of the observed variables. The constraint-based approach does not have to rely on the assumption that there are no common confounders, and, as a result, can sometimes detect the presence of confounders between observed variables from the data (Spirtes et al., 2000). A drawback of this approach is lack of robustness in some cases. Typical implementations makes use of independence tests, making the results for borderline independencies/dependencies incorrect sometimes (Spirtes et al., 2000). As a result, the outcome of learning a network can be sensitive to such errors. In this study we applied a state-of-the-art algorithm for structure learning called Bayesian Constraint-based Causal Discovery (BCCD) (Claassen and Heskes, 2012a) to infer the causal structure from the data. BCCD combines the strength of constraint-based and score-based approaches, which allows it to outperform the best algorithms in the field (Claassen and Heskes, 2012a)(Claassen, Heskes, 2012). This algorithm is able to detect common causes of the observed variables similar to constraint-based approaches and provides a reliability measure of the inferred relationship like the score-based approach. This reliability measure gives a conservative estimate of the probability of a causal relation. A recently extended version of BCCD can handle data that contains a mixture of discrete and continuous variables and missing values and does not require discretization or imputation that can lead to loss of information or biased results as described in Chapter 2.

BCCD can handle directed acyclic graphs that contain latent variables. These graphs are called maximal ancestral graphs (MAG). All MAGs that represent the same set of conditional independencies form an equivalence class. The equiv-
A change class for MAGs is called a partial ancestral graph (PAG). The BCCD algorithm produces PAGs as an output. An edge between two variables in PAG suggests that there is a direct causal relationship between them. This can be either an effect of one variable on another (“A → B”), unobserved common cause “A ↔ B” or a selection bias “A ← B”. If a direction of an edge between two variables is non-identifiable it is marked with a circle mark “◦”. No edge between variables means that these variables are conditionally independent given other variables in the network. For example, if A and B are correlated but there is no edge between them in the PAG, that implies that there is either an indirect causal path from one variable to another through some other variables in this PAG, a common cause between these variables, or a selection bias.

6.5.2 Results BCCD

Running the BCCD algorithm provided three tables (Appendix Tables 6.2, 6.3, and 6.4). Table 6.2 provides the reliability of the statement that a direct link exists between two variables. Table 6.3 presents the reliability of the causal statement: “A causes B”, both for direct and indirect causal effects. If Table 6.3 says that “A causes B” and there is an edge between A and B, this edge has a tail from A to B and A causes B in the PAG. For example, the variable “Gender” caused variable “Inattention” with reliability 51%. Table 6.4 represents the reliability of the causal statement: “A does not cause B”, both for direct and indirect causal effects. If Table 6.4 says that “A does not cause B” and there is an edge between A and B, this edge has an arrow head from B to A. For example, the variable “PIQ” does not cause variable “VIQ” with reliability 53%. The prior knowledge used in the model, e.g. that gender is not caused by other variables and, is represented in Table 6.4 in cells with a reliability of 100/

The difference between Tables 6.3 and 6.4 compared to Table 6.2 is that Tables 6.3 and 6.4 give an estimate for the direction of the causal effect, whereas Table 6.2 provides estimates for the presence of the edge between two variables. Moreover, Tables 6.3 and 6.4 show the reliability of the direction of both direct and indirect causal paths, whereas Table 6.2 gives a reliability estimate for a direct causal path between two variables only.

6.5.3 Mediation analysis

We applied mediation analysis to investigate whether there is a direct link between hyperactivity trait of ADHD and social ineptness trait of ASD. We built a regression model where social ineptness trait is a dependent variable, inattention and impulsivity are possible mediators and hyperactivity is an independent
Table 6.2: Reliability of direct links between two variables.

<table>
<thead>
<tr>
<th>Age</th>
<th>Gender</th>
<th>Reduced contact</th>
<th>Social ineptness</th>
<th>Repetitive behavior</th>
<th>Fear of changes</th>
<th>Verbal IQ</th>
<th>Performance IQ</th>
<th>Inattention</th>
<th>Hyperactivity</th>
<th>Impulsivity</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.99</td>
<td>0.98</td>
<td>0.88</td>
<td>0.88</td>
<td>0.99</td>
<td>0.99</td>
<td>0.99</td>
<td>0.99</td>
<td>0.99</td>
<td>0.99</td>
<td>0.99</td>
</tr>
<tr>
<td>0.97</td>
<td>0.96</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>0.95</td>
<td>0.94</td>
<td>0.93</td>
<td>0.93</td>
<td>0.93</td>
<td>0.93</td>
<td>0.93</td>
<td>0.93</td>
<td>0.93</td>
<td>0.93</td>
<td>0.93</td>
</tr>
</tbody>
</table>

Table 6.3: The reliability estimate of the logical statement “A causes B”, where A is represented in rows and B in columns. The estimate is provided for logical statements with reliability of 50% or higher.

<table>
<thead>
<tr>
<th>Age</th>
<th>Gender</th>
<th>Reduced contact</th>
<th>Social ineptness</th>
<th>Repetitive behavior</th>
<th>Fear of changes</th>
<th>Verbal IQ</th>
<th>Performance IQ</th>
<th>Inattention</th>
<th>Hyperactivity</th>
<th>Impulsivity</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.99</td>
<td>0.98</td>
<td>0.98</td>
<td>0.98</td>
<td>0.98</td>
<td>0.98</td>
<td>0.98</td>
<td>0.98</td>
<td>0.98</td>
<td>0.98</td>
<td>0.98</td>
</tr>
<tr>
<td>0.97</td>
<td>0.96</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>0.95</td>
<td>0.94</td>
<td>0.93</td>
<td>0.93</td>
<td>0.93</td>
<td>0.93</td>
<td>0.93</td>
<td>0.93</td>
<td>0.93</td>
<td>0.93</td>
<td>0.93</td>
</tr>
</tbody>
</table>

Table 6.4: The reliability of the logical statement “A does not cause B”, where A is represented in rows and B in columns. The estimate is provided for logical statements with reliability of 50% or higher. Cells with a reliability of 100% represent the prior knowledge used in the model, e.g. that gender is not caused by other variables.

<table>
<thead>
<tr>
<th>Age</th>
<th>Gender</th>
<th>Reduced contact</th>
<th>Social ineptness</th>
<th>Repetitive behavior</th>
<th>Fear of changes</th>
<th>Verbal IQ</th>
<th>Performance IQ</th>
<th>Inattention</th>
<th>Hyperactivity</th>
<th>Impulsivity</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.99</td>
<td>0.98</td>
<td>0.98</td>
<td>0.98</td>
<td>0.98</td>
<td>0.98</td>
<td>0.98</td>
<td>0.98</td>
<td>0.98</td>
<td>0.98</td>
<td>0.98</td>
</tr>
<tr>
<td>0.97</td>
<td>0.96</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>0.95</td>
<td>0.94</td>
<td>0.93</td>
<td>0.93</td>
<td>0.93</td>
<td>0.93</td>
<td>0.93</td>
<td>0.93</td>
<td>0.93</td>
<td>0.93</td>
<td>0.93</td>
</tr>
</tbody>
</table>

110
variables. Regression analysis showed that hyperactivity is not a significant predictor of social ineptness ($\beta = 0.01, p = 0.79$). Thus, we can conclude that there is no direct link between hyperactivity and social ineptness.

Using mediation analysis we tested whether there is an evidence of no direct link between inattention and repetitive behavior as well as between impulsivity and repetitive behavior. We built a regression model where repetitive behavior is a dependent variable, social ineptness, and hyperactivity are possible mediators, and inattention (Figure 6.3) or impulsivity (Figure S1b) are independent variables. Analysis showed that the regression coefficients are not significant between inattention and repetitive behavior ($\beta = 0.03, p = 0.11$), and between impulsivity and repetitive behavior ($\beta = 0.01, p = 0.86$).

We investigated whether the direct and indirect associations between verbal IQ with ADHD and ASD symptoms described above can be explained with mediation analysis. First we explored whether verbal IQ is associated directly only with one ADHD trait (inattention). To check this hypothesis, we built a regression model, where verbal IQ is a dependent variable, inattention is a possible
mediator, and hyperactivity (Figure 6.4a), and impulsivity (Figure 6.4b) are independent variables. Our analysis showed that neither hyperactivity (β = −0.25, p = 0.16) nor for impulsivity (β = 0.09, p = 0.74) are significant predictor for verbal IQ when inattention trait is present in the regression model. This confirms that the association between verbal IQ and impulsivity as well as hyperactivity is not direct and mediated through inattention.

Then we investigated whether verbal IQ is not directly associated with repetitive behavior, reduced contact and fear of changes trait of ASD as was shown by BCCD. We built a regression model where verbal IQ is a dependent variable, social ineptness and hyperactivity (only for model with repetitive behavior) are possible mediators, and repetitive behavior (Figure 6.5a), reduced contact (Figure 6.5b), and fear of changes (Figure 6.5c) are an independent variable. Mediation analysis showed that the regression coefficients of repetitive behavior (β = −0.27, p = 0.12), reduced contact (β = 0.02, p = 0.87) and fear of changes (β = 0.04, p = 0.89) are not significant in such a model, suggesting that there is no direct effect between these variables and verbal IQ and, but this effect is mediated via the other traits. This confirms our previous findings with the BCCD algorithm.
Figure 6.5: Regression model for mediation analysis that predicts dependent variable (in grey) verbal IQ using social ineptness (and hyperactivity in (a)) as a mediator and repetitive behavior (a), reduced contact (b), or fear of changes (c) as independent predictor. The regression model is presented at the top of the figure, the significance of the regression coefficient is shown next to the edge.
Figure 6.6: Output causal model representing causal relationships between variables in the ADHD ASD data set when reduced to one subject per family. Edge directions are marked with ‘-’ and ‘>’ for identifiable edge directions and with ‘o’ for non-identifiable edge directions. Reliability estimates for the presence of an edge are depicted as percentage. Direct links between ASD, ADHD and IQ are marked in red.
Conclusions

The main questions that we wanted to answer in this thesis were: a) how to make causal discovery algorithms more general so that they could be applied in a larger spectrum of problems; b) can we explain interactions of different factors with ADHD with the help of causal discovery.

To tackle the first question we extended existing causal discovery algorithms to be able to handle both discrete and continuous variables and missing data. The proposed method is suitable for a broad range of real-world data sets, since it only requires data to have non-paranormal distribution and missing values missing at random. The limitation of the method is a restriction to monotonic dependencies between variables, and small number of variables due to a computational complexity. Working only with small data sets (10-25 variables) restricts the set of possible applications of the method. Although smaller causal models are easier to interpret, the current trends in data science requires the algorithms to be able to handle data sets with thousands of variables and millions of rows. As a future work, we would like to extend this method to work with large data sets and more complex dependencies.

In Chapter 3 we targeted another challenge that arises when applying causal modeling, the estimation of the reliability of this model. Although in our study our approach was integrated in the BCCD algorithm, in principle it can be applied to any constraint-based causal discovery algorithm that infers causal statements. Thus, as a next step we would like to integrate this method in other constraint-based algorithms and evaluate the improvement of the model reliability estimate. The proposed method uses Fréchet inequalities to estimate the lower and upper bounds of causal statements, however, other approaches such as linear programming can also be used to compute the bounds. As future work, we would like to use SAT solvers to provide a more accurate estimate of
the lower and upper bound than using Fréchet inequalities.

To answer the second question of our study we applied causal modeling to seven different data sets about ADHD. The results of this modeling provide new insights into the association of genetic factors with ADHD symptoms, behavior problems, and comorbid disorders. Although most patients with ADHD have a combined type where both inattention and hyperactive/impulsive symptoms are present, our study showed that these two symptom domains interact differently with ADHD risk factors and behavior problems. Causal modeling showed an importance of inattention as a leading symptom within the disorder. To confirm the conclusions of the inferred models, experiments should be performed. As a next step in the study of ADHD we would like to extend the number of comorbid disorders and possible genetic factors in one model to provide a global causal model of ADHD. Moreover, we would like to explain the evolution of ADHD with aging of the patient, using causal modeling. That will help to understand how the treatment should be adjusted based on patient’s age.

The results of the causal modeling suggest that although ADHD symptom domains are highly correlated the interactions with the possible risk factors, comorbid disorders, and behavior status is different. Strong correlation between symptom domains can lead to the problem that correlation analysis will show that all domains are associated with a variable of interest; however, in reality the direct interaction is present only with one symptom domain as we have shown in Chapters 2-6. This shows the importance of causal modeling to infer the causal relationship between variables and reduce significantly the number of potential hypothesis to test on practice.

To demonstrate the difference between BCCD and more standard causal discovery approaches, we applied the PC algorithm to the data set discussed in Chapter 4. To estimate the reliability of the presence of an edge in the model, we bootstrapped the data 150 times and rerun the algorithm. As a result, we inferred a model shown in Figure 7.1. We show only the edges with the reliability higher than 50%. This model has links between DAT1 haplotype and inattention, between symptoms and patient/control variable, and between patient/control and medication. However, this model does not have links from patient/control to striatal activation and medication (see Figure 4.1). This suggests that the BCCD algorithm is more sensitive that the standard PC algorithm. Moreover, the standard PC algorithm does not provide the reliability of edge directions.

In recent years the application of advanced statistical models to analyze medical data and the etiology of ADHD increased significantly. That provides an opportunity to compare the results obtained in this thesis with other studies in the same topic. In paper (Pingault et al., 2015) the authors examined the genetic and environmental influences of ADHD symptoms using a latent growth
curve model. The results of the study suggested significant difference in the genetic effect on inattention domain vs. hyperactivity/impulsivity domain. In particular, study showed large dominant and additive genetic effects for inattention symptoms and only additive genetic effects for hyperactive/impulsive symptoms. These results are in line with the causal model for the interaction of inattention and hyperactivity/impulsivity and causal factors obtained in Chapter 4.

In another paper (de Zeeuw et al., 2017) the authors try to answer questions of causal relationships between ADHD and low educational achievement, having a similar line of argumentation as the one described in Chapter 4. However, this paper provides an alternative approach to detect causal relationships in comparison to Chapter 4. The limitation of this method is that it requires longitudinal data from a twin study. The results of the study confirmed the initial hypothesis of direct effect of ADHD on the level of education achievement. Moreover, the results were also supported by the analysis of the effect of the
medication on educational achievement. These findings explain strong association between IQ level and ADHD measured in all causal models in this study, and suggest that this effect goes from ADHD to IQ. This agrees with the effect of ADHD on IQ found in the causal models in Chapter 2 and 3.

In the study (Knouse et al., 2013) the authors employed SEM to examine the effect of the disorder on the personality traits. This study suggested a different association between ADHD symptom trait and personality traits in particular inattention was positively associated with neuroticism and negatively associated with conscientiousness. On the other hand, hyperactivity and impulsivity showed differential relationships to extraversion and agreeableness. Moreover, hyperactivity was positively associated with conscientiousness. The SEM inferred in (Knouse et al., 2013) can be used to extend the models provided in this thesis to describe the causal relationships between different variables within ADHD.

The emergence of these studies shows that the application of causal modeling to data can improve the understanding of the etiology of ADHD and accelerate research in this area.
Summary

This thesis is focusing on the application of causal discovery in the medical domain. The first goal of this thesis was to extend the existing algorithms in the causal discovery domain in order to relax unrealistic assumptions that current methods rely on. In Chapter 2 we relaxed two main assumptions: data either obeys a Gaussian distribution or is discrete, and data has no missing values. In the medical domain these assumptions often do not hold, since data can have both discrete and continuous data, where continuous data is not necessary Gaussian. We developed a method that can infer the structure of a causal graph when the data set contains both discrete and continuous variables and also has missing values. We proposed several approaches to solve this problem and based on a simulation study concluded that the approach that uses the expectation maximization algorithm leads to the best performance. The advantage of this method is also that it assumes that data is missing at random, while alternative approaches assume that the data is missing completely at random, which is a much stronger assumption.

In Chapter 3 we continued extending existing causal discovery algorithms and considered the problem of estimating the reliability of causal statements. The reliability of the causal statement can be easily estimated for base statements that are inferred based on conditional (in)dependencies. However, these statements represent only a part of the causal statements that are used to infer causal relationships between variables. The other part is estimated by applying relatively straightforward causal logic such as transitivity and acyclicity to base statements. In Chapter 3 we proposed a new approach to estimate the reliability of novel statements inferred by forward chaining. In order to do that, we applied Fréchet inequalities and logic of predicates. The main challenge of this approach is computational infeasibility of calculations for large graphs. To
overcome this problem, we proposed two versions of the algorithm that provide different accuracies of the final reliability estimate: a greedy and a delayed version. Simulation studies showed that the delayed variant, at the expense of more bookkeeping and computation time, does provide slightly tighter intervals.

The second goal of the thesis was the application of the developed methods to real world data sets describing ADHD. With the help of causal discovery our aim was to provide a better understanding of the etiology of ADHD, to develop a large model that can include different potential risk factors, and understand the interactions of ADHD with other disorders that often co-occur with ADHD, such as aggression and ASD. In Chapters 2 and 3 we inferred three models that describe ADHD, some of the risk factors (gender, genes, age, etc.), behavioral information, and measurements of activations in particular brain regions. These results provide new insight into the reversal learning problem and can improve its treatment. Moreover, the causal model suggests promising new pathways for genetic research in ADHD that need to be confirmed by genetic imaging studies.

In Chapter 4-6 the focus is on answering questions about the etiology of ADHD using causal modeling as a tool. In Chapter 4 we used the algorithm from Chapter 2 to determine whether there is a direct or indirect link between a candidate gene for ADHD and brain activation during a monetary incentive delay task. The causal model inferred from the data showed that there is no direct link between these two variables; however, there is an indirect path between them. This finding might explain existing discrepancies in the current literature, but to make solid conclusions further experiments should be performed.

In Chapter 5 we applied causal modeling to answer the question whether inattention drives hyperactivity/impulsivity or hyperactivity/impulsivity drives inattention in ADHD. Based on the results of the causal modeling the most probable model explaining the conditional independencies observed in the data is the model where inattention causes hyperactivity/impulsivity. We discuss our results providing a historical context and relate our results with other studies that showed different etiology of the hyperactivity/impulsivity for subjects that have a high level of inattention from subjects with a low level of inattention. The hypothesis proposed in this chapter can be further validated experimentally.

In Chapter 6 we studied the interactions between ADHD and its comorbid disorder ASD. The causal modeling builds a more complete model than standard statistical techniques, distinguishes between direct and indirect associations, and allows us to make preliminary predictions about causation. Applying causal modeling to a large phenotypic data set suggested three distinct pathways between the disorders where the strongest link was found between social communication difficulties, inattention, and impulsivity. These findings may help
future studies on understanding the (pathophysiological) mechanisms behind
the overlap between ASD and ADHD.

In conclusion, this thesis provides a great tool to infer causal relationships
from observational data and provides several examples of its application to real
world medical data. Based on the inferred models new hypothesis for experi-
mental studies are proposed that can help scientists to understand the mecha-
nisms of ADHD.
Samenvatting

Dit proefschrift richt zich op het leren van causale modellen op basis van medische data. Het eerste doel van dit proefschrift was het aanpassen van bestaande causale methoden zodat deze minder strikte aannamen hoeven te maken. In hoofdstuk 2 verzwakten we twee aannames: data volgt een normale of discrete verdeling, en data heeft geen ontbrekende waarden. In het medische domein zijn deze aannamen vaak niet waar, aangezien data zowel discrete als continue data kan bevatten, waarbij continue data niet noodzakelijk normaal verdeeld is. We ontwikkelden een methode die de structuur van een causaal model kan afleiden wanneer de dataset zowel discrete als continue variabelen bevat en er mogelijke waarden ontbreken. We hebben verschillende benaderingen voorgesteld om dit probleem op te lossen en op basis van een simulatie studie is geconcludeerd dat de aanpak die het expectation maximization algoritme gebruikt, tot de beste prestatie leidt. Het voordelen van deze methode is ook dat men ervan uitgaat dat data willekeurig ontbreekt (missing at random, MAR), terwijl alternatieve benaderingen ervan uitgaan dat de data volledig willekeurig ontbreekt (missing completely at random MCAR), wat een veel sterker aanname is.

In Hoofdstuk 3 breiden we bestaande causale algoritmen uit en beschouwen we het probleem van het schatten van de betrouwbaarheid van causale relaties. De betrouwbaarheid van een causale relatie die afgeleid is van voorwaardelijke afhankelijkheden kan gemakkelijk worden geschat. Dit noemen we de basis causale uitspraken. Deze uitspraken vertegenwoordigen echter slechts een deel van de causale uitspraken die gebruikt worden om causale relaties tussen variabelen af te leiden. Voor de andere causale uitspraken wordt de betrouwbaarheid geschat door relatief eenvoudige causale logica toe te passen, zoals transitiviteit en acycliciteit van basis uitspraken. In hoofdstuk 3 stelden we een nieuwe aanpak voor om de betrouwbaarheid van nieuwe afgeleide uitspraken
te berekenen. Om dit te doen pasten we Fréchet ongelijkheden en predikaatlogica toe. De belangrijkste uitdaging van deze aanpak is de complexiteit van de vele berekeningen voor grote modellen. Om dit probleem te verhelpen, stelden we twee versies van het algoritme voor die de betrouwbaarheid van causale uitspraken met verschillende nauwkeurigheden kunnen bepalen: een gulzige en een vertraagde versie. Simulatiestudies toonden aan dat de vertraagde variant, ten koste van meer boekhoud- en rekentijd, iets nauwkeurigere intervallen berekent.

Het tweede doel van het proefschrift was het toepassen van de ontwikkelde methoden op echte datasets die ADHD beschrijven. Met behulp van causale modellen streefden we ernaar om beter inzicht te krijgen in de etiologie van ADHD. Daarvoor wilden we een groot model ontwikkelen dat verschillende mogelijke risicofactoren en de interacties van ADHD met andere aandoeningen die vaak voorkomen bij ADHD, zoals agressie en ASD kan beschrijven. In hoofdstukken 2 en 3 ontwikkelden we drie modellen die ADHD, een aantal risicofactoren (geslacht, genen, leeftijd, enz.), gedragsinformatie en metingen van activaties in specifieke hersengebieden bevatten. Deze resultaten geven een nieuw inzicht in het omgekeerde leerprobleem en kunnen gebruikt worden om de behandeling van ADHH te verbeteren. Bovendien stelt het causaal model veelbelovende nieuwe trajecten van genetisch onderzoek in ADHD voor, die met behulp van genetische studies kunnen worden bevestigd.

In hoofdstuk 4-6 ligt de nadruk op het beantwoorden van vragen over de etiologie van ADHD met behulp van causale modellering als hulpmiddel. In hoofdstuk 4 gebruikten we het algoritme uit hoofdstuk 2 om te bepalen of er een directe of indirecte verbinding bestaat tussen een kandidaatgen voor ADHD en hersenactivering tijdens een monetaire stimulansvertraging. Het causaal model dat uit de data is afgeleid, laat zien dat er geen directe verbinding bestaat tussen deze twee variabelen, maar wel dat er een indirect pad bestaat. Deze bevinding zou de bestaande discrepanties in de huidige literatuur kunnen verklaren, maar om solide conclusies te maken, moeten er experimenten worden uitgevoerd.

In hoofdstuk 5 pasten we causale modellering toe om de vraag te beantwoorden of innatentie hyperactiviteit / impulsiviteit aandrijft of hyperactiviteit / impulsiviteit innatentie aandrijft in ADHD. Op basis van de resultaten van de causale modellering is het meest waarschijnlijke model waarin de voorwaardelijke onafhankelijkheden in de data worden waargenomen, het model waarbij inattentie hyperactiviteit / impulsiviteit veroorzaakt. We bespreken onze resultaten in een historische context en vergelijken onze resultaten met andere studies die verschillende etiologie van de hyperactiviteit / impulsiviteit toonden voor onderwerpen die een hoog niveau van inattentie hebben van onderwerpen met een laag niveau van inattentie. De hypothese die in dit hoofdstuk wordt
voorgesteld, kan experimenteel worden gevalideerd.

In hoofdstuk 6 hebben we de interacties tussen ADHD en zijn comorbid disorde ASD bestudeerd. De causale modellering geeft een model dat naukeuriger is dan standaard statistische technieken, onderscheidt directe en indirecte associaties en stelt ons in staat om voorspellingen te doen over causale verbanden. Toepassing van causale modellering op een grote fenotypische dataset stelde drie verschillende verbindingen tussen de aandoeningen voor. De sterkste verbinding werd gevonden tussen sociale communicatieproblemen, innatentie en impulsiviteit. Deze bevindingen kunnen toekomstige studies helpen bij het begrijpen van de (pathofysiologische) mechanismen achter de overlap tussen ASD en ADHD.

Ten slotte biedt dit proefschrift een goed instrument om causale relaties uit observaties af te leiden en geeft een aantal voorbeelden van de toepassing ervan op medische data. Op basis van de afgeleide modellen worden nieuwe hypothesen voor experimentele studies voorgesteld die de wetenschappers kunnen helpen om de mechanismen van ADHD beter te begrijpen.
Curriculum Vitae

Elena was born in Novosibirsk, Russia. She obtained her Master’s Degree in Applied Mathematics in 2008 in Novosibirsk State University. In 2010 she came to the Netherlands to study in the postgraduate program industrial mathematics at the Technical Eindhoven University (TU/e). She graduated from the TU/e in 2012 with a diploma of professional doctorate in engineering.

Elena started her PhD in Data Science group at the Institute for Computing and Information Sciences at the Radboud University in 2012 under the supervision of prof. dr. Tom Heskes, prof. dr. Jan Buitelaar, dr. Perry Groot, and dr. Jeffrey Glennon. During her PhD she supervised master students, assisted in several courses (machine Learning in practice, data mining, and information retrieval), did an internship as data scientist at Philips Research, and presented her work at several international conferences.

Currently Elena is working as a data scientist at Booking.com in Amsterdam, where she works on building machine learning models to personalize the website for different user needs. She runs her models in production on real traffic and tests them by means of experimentation and A/B testing.
Chapter 11

Acknowledgments

As the last part of this thesis I would like to say thanks to everyone who worked with me in these four years, helped me, supported me, and added some fun in our university life. Each of you changed my life and taught me something new. Whether it was something about research, machine learning, different cultures, running, or baking cakes, I appreciated that a lot! I would like to thank all my colleagues: Elena M, Joris, Daniel, Wout, Alexandra, Daniela, Wessel, Fabian, Saiden, Tameem, Suzan, Herman, Arjen, Theo, Marieke, Saskia, Maya, Max, Jacopo, Twan, Robbert, Henning, Jonce, Tom C, Ruifei, Gido, Mohsen, Christiaan, Tjeerd, Johannes, Simone, Gabriel, Nicole, Kasper, and Ridho. Thank you for all the fun that we had together, for our great Sinterklaas parties, for debates during lunch, for coffee breaks every morning and of course for the cake competition that made our life a bit sweeter!

I want to thank all my colleagues from the medical center that helped me to learn so much in the domain of ADHD and ASD. Thank you Daniel, Catarina, Nanda, Barbara, Martine, Jill and Anoek! I also want to thank Helen and all my colleagues at Philips for giving me a chance to get some industry experience and contribute to ACT project.

I really want to thank my supervisors who supported me, guided me in this difficult path of learning how to be a researcher. Thank you Jeffrey, Perry, Jan! And I would like to say a special thanks to Tom. For the open door policy, for being open to different opinions, for giving freedom and for appreciation of initiative. Thanks for surprising me so many times. When I thought that our problem was unsolvable, you came with a solution, or when I thought that my solution was perfect, you spotted its drawbacks. It taught me to be more precise and more critical to my work. And special thanks for being a professor that does not endorse hierarchy but treats a student as an equal colleague and gives
a chance to disagree and support your point of view.

I would like to also thank those people who motivated me to become a mathematician throughout my life. My teacher of Math at school Ludmila Igorevna Bulgakova. All my teachers from NSU and especially my teacher of mathematical analysis Kapranov Jurij Ivanovich. My teacher of probability and statistics Natalia Isaakovna Chernova. Her lessons made me really excited about probability theory and motivated me to do a PhD in machine learning. My supervisors at the TU/e Stef van Eijndhoven and Cor Hurkens, and my teacher of applied statistics Koo Rijpkema. And again thanks to everyone in Nijmegen who taught me so much about causality.

I also want to thank all my family, my father, sister, my husband for their support, genes of mathematicians, for believing in me and helping me to achieve my goals and for all your love. And the last I want to thank my mother. She is the person who made me a true mathematician. Starting making math exercises with me at the age of 5, at the age of 11 I was already solving at home tasks from entry exams from the math department at the university under her supervision. She made me interested in math, guided me, supported me and always believed that I am able to become a researcher. You are the main teacher in my life and I dedicate this book to you, mama.
SIKS dissertations

2011 01 Botond Cseke (RUN), Variational Algorithms for Bayesian Inference in Latent Gaussian Models
02 Nick Tinnemeier (UU), Organizing Agent Organizations. Syntax and Operational Semantics of an Organization-Oriented Programming Language
03 Jan Martijn van der Werf (TUE), Compositional Design and Verification of Component-Based Information Systems
04 Hado van Hasselt (UU), Insights in Reinforcement Learning; Formal analysis and empirical evaluation of temporal-difference
05 Bas van der Raadt (VU), Enterprise Architecture Coming of Age - Increasing the Performance of an Emerging Discipline.
06 Yiwen Wang (TUE), Semantically-Enhanced Recommendations in Cultural Heritage
07 Yujia Cao (UT), Multimodal Information Presentation for High Load Human Computer Interaction
08 Nieske Vergunst (UU), BDI-based Generation of Robust Task-Oriented Dialogues
09 Tim de Jong (OU), Contextualised Mobile Media for Learning
10 Bart Bogaert (UvT), Cloud Content Contention
11 Dhaval Vyas (UT), Designing for Awareness: An Experience-focused HCI Perspective
12 Carmen Bratosin (TUE), Grid Architecture for Distributed Process Mining
13 Xiaoyu Mao (UvT), Airport under Control. Multiagent Scheduling for Airport Ground Handling
14 Milan Lovric (EUR), Behavioral Finance and Agent-Based Artificial Markets
Chapter 12. SIKS dissertations

15 Marijn Koolen (UvA), The Meaning of Structure: the Value of Link Evidence for Information Retrieval
16 Maarten Schadd (UM), Selective Search in Games of Different Complexity
17 Jiyin He (UVA), Exploring Topic Structure: Coherence, Diversity and Relatedness
18 Mark Ponsen (UM), Strategic Decision-Making in complex games
19 Ellen Rusman (OU), The Mind’s Eye on Personal Profiles
20 Qing Gu (VU), Guiding service-oriented software engineering - A view-based approach
21 Linda Terlouw (TUD), Modularization and Specification of Service-Oriented Systems
22 Junte Zhang (UVA), System Evaluation of Archival Description and Access
23 Wouter Weerkamp (UVA), Finding People and their Utterances in Social Media
24 Herwin van Welbergen (UT), Behavior Generation for Interpersonal Coordination with Virtual Humans On Specifying, Scheduling and Realizing Multimodal Virtual Human Behavior
25 Syed Waqar ul Qounain Jaffry (VU), Analysis and Validation of Models for Trust Dynamics
26 Matthijs Aart Pontier (VU), Virtual Agents for Human Communication - Emotion Regulation and Involvement-Distance Trade-Offs in Embodied Conversational Agents and Robots
27 Aniel Bhulai (VU), Dynamic website optimization through autonomous management of design patterns
28 Rianne Kaptein (UVA), Effective Focused Retrieval by Exploiting Query Context and Document Structure
29 Faisal Kamiran (TUE), Discrimination-aware Classification
30 Egon van den Broek (UT), Affective Signal Processing (ASP): Unraveling the mystery of emotions
31 Ludo Waltman (EUR), Computational and Game-Theoretic Approaches for Modeling Bounded Rationality
32 Nees-Jan van Eck (EUR), Methodological Advances in Bibliometric Mapping of Science
33 Tom van der Weide (UU), Arguing to Motivate Decisions
34 Paolo Turrini (UU), Strategic Reasoning in Interdependence: Logical and Game-theoretical Investigations
35 Maaike Harbers (UU), Explaining Agent Behavior in Virtual Training
36 Erik van der Spek (UU), Experiments in serious game design: a cognitive approach

132
Adriana Burlutiu (RUN), Machine Learning for Pairwise Data, Applications for Preference Learning and Supervised Network Inference

Nyree Lemmens (UM), Bee-inspired Distributed Optimization

Joost Westra (UU), Organizing Adaptation using Agents in Serious Games

Viktor Clerc (VU), Architectural Knowledge Management in Global Software Development

Luan Ibraimi (UT), Cryptographically Enforced Distributed Data Access Control

Michal Sindlar (UU), Explaining Behavior through Mental State Attribution

Henk van der Schuur (UU), Process Improvement through Software Operation Knowledge

Boris Reuderink (UT), Robust Brain-Computer Interfaces

Herman Stehouwer (UvT), Statistical Language Models for Alternative Sequence Selection

Beibei Hu (TUD), Towards Contextualized Information Delivery: A Rule-based Architecture for the Domain of Mobile Police Work

Azizi Bin Ab Aziz (VU), Exploring Computational Models for Intelligent Support of Persons with Depression

Mark Ter Maat (UT), Response Selection and Turn-taking for a Sensitive Artificial Listening Agent

Andreea Niculescu (UT), Conversational interfaces for task-oriented spoken dialogues: design aspects influencing interaction quality

Terry Kakeeto (UvT), Relationship Marketing for SMEs in Uganda

Muhammad Umair (VU), Adaptivity, emotion, and Rationality in Human and Ambient Agent Models

Adam Vanya (VU), Supporting Architecture Evolution by Mining Software Repositories

Jurriaan Souer (UU), Development of Content Management System-based Web Applications

Marijn Plomp (UU), Maturing Interorganisational Information Systems

Wolfgang Reinhardt (OU), Awareness Support for Knowledge Workers in Research Networks

Rianne van Lambalgen (VU), When the Going Gets Tough: Exploring Agent-based Models of Human Performance under Demanding Conditions

Gerben de Vries (UVA), Kernel Methods for Vessel Trajectories

Ricardo Neisse (UT), Trust and Privacy Management Support for Context-Aware Service Platforms

David Smits (TUE), Towards a Generic Distributed Adaptive Hypermedia Environment
11 J.C.B. Rantham Prabhakara (TUE), Process Mining in the Large: Pre-processing, Discovery, and Diagnostics
12 Kees van der Sluijs (TUE), Model Driven Design and Data Integration in Semantic Web Information Systems
13 Suleman Shahid (UvT), Fun and Face: Exploring non-verbal expressions of emotion during playful interactions
14 Evgeny Knutov (TUE), Generic Adaptation Framework for Unifying Adaptive Web-based Systems
15 Natalie van der Wal (VU), Social Agents. Agent-Based Modelling of Integrated Internal and Social Dynamics of Cognitive and Affective Processes.
16 Fiemke Both (VU), Helping people by understanding them - Ambient Agents supporting task execution and depression treatment
17 Amal Elgammal (UvT), Towards a Comprehensive Framework for Business Process Compliance
18 Eltjo Poort (VU), Improving Solution Architecting Practices
19 Helen Schonenberg (TUE), What’s Next? Operational Support for Business Process Execution
20 Ali Bahramisharif (RUN), Covert Visual Spatial Attention, a Robust Paradigm for Brain-Computer Interfacing
21 Roberto Cornacchia (TUD), Querying Sparse Matrices for Information Retrieval
22 Thijs Vis (UvT), Intelligence, politie en veiligheidsdienst: verenigbare grootheden?
23 Christian Muehl (UT), Toward Affective Brain-Computer Interfaces: Exploring the Neurophysiology of Affect during Human Media Interaction
24 Laurens van der Werff (UT), Evaluation of Noisy Transcripts for Spoken Document Retrieval
25 Silja Eckartz (UT), Managing the Business Case Development in Inter-Organizational IT Projects: A Methodology and its Application
26 Emile de Maat (UVA), Making Sense of Legal Text
27 Hayrettin Gurkok (UT), Mind the Sheep! User Experience Evaluation & Brain-Computer Interface Games
28 Nancy Pascall (UvT), Engendering Technology Empowering Women
29 Almer Tigelaar (UT), Peer-to-Peer Information Retrieval
30 Alina Pommeranz (TUD), Designing Human-Centered Systems for Reflective Decision Making
31 Emily Bagarukayo (RUN), A Learning by Construction Approach for Higher Order Cognitive Skills Improvement, Building Capacity and Infrastructure
32 Wietske Visser (TUD), Qualitative multi-criteria preference representation and reasoning
33 Rory Sie (OUN), Coalitions in Cooperation Networks (COCOON)
34 Pavol Jancura (RUN), Evolutionary analysis in PPI networks and applications
35 Evert Haasdijk (VU), Never Too Old To Learn – On-line Evolution of Controllers in Swarm- and Modular Robotics
36 Denis Ssebugwawo (RUN), Analysis and Evaluation of Collaborative Modeling Processes
37 Agnes Nakakawa (RUN), A Collaboration Process for Enterprise Architecture Creation
38 Selmar Smit (VU), Parameter Tuning and Scientific Testing in Evolutionary Algorithms
39 Hassan Fatemi (UT), Risk-aware design of value and coordination networks
40 Agus Gunawan (UvT), Information Access for SMEs in Indonesia
41 Sebastian Kelle (OU), Game Design Patterns for Learning
42 Dominique Verpoorten (OU), Reflection Amplifiers in self-regulated Learning
43 Withdrawn
44 Anna Tordai (VU), On Combining Alignment Techniques
45 Benedikt Kratz (UvT), A Model and Language for Business-aware Transactions
46 Simon Carter (UVA), Exploration and Exploitation of Multilingual Data for Statistical Machine Translation
47 Manos Tsagkias (UVA), Mining Social Media: Tracking Content and Predicting Behavior
48 Jorn Bakker (TUE), Handling Abrupt Changes in Evolving Time-series Data
49 Michael Kaisers (UM), Learning against Learning - Evolutionary dynamics of reinforcement learning algorithms in strategic interactions
50 Steven van Kervel (TUD), Ontology driven Enterprise Information Systems Engineering
51 Jeroen de Jong (TUD), Heuristics in Dynamic Scheduling; a practical framework with a case study in elevator dispatching

2013 01 Viorel Milea (EUR), News Analytics for Financial Decision Support
02 Erietta Liarou (CWI), MonetDB/DataCell: Leveraging the Column-store Database Technology for Efficient and Scalable Stream Processing
03 Szymon Klarman (VU), Reasoning with Contexts in Description Logics
04 Chetan Yadati (TUD), Coordinating autonomous planning and scheduling
05 Dulce Pumareja (UT), Groupware Requirements Evolutions Patterns
06 Romulo Goncalves (CWI), The Data Cyclotron: Juggling Data and Queries for a Data Warehouse Audience
07 Giel van Lankveld (UvT), Quantifying Individual Player Differences
08 Robbert-Jan Merk (VU), Making enemies: cognitive modeling for opponent agents in fighter pilot simulators
09 Fabio Gori (RUN), Metagenomic Data Analysis: Computational Methods and Applications
10 Jeewanie Jayasinghe Arachchige (UvT), A Unified Modeling Framework for Service Design.
11 Evangelos Pournaras (TUD), Multi-level Reconfigurable Self-organization in Overlay Services
12 Mohammad Safiri (UT), Service Tailoring: User-centric creation of integrated IT-based homecare services to support independent living of elderly
13 Jafar Tanha (UVA), Ensemble Approaches to Semi-Supervised Learning
14 Daniel Hennes (UM), Multiagent Learning - Dynamic Games and Applications
15 Eric Kok (UU), Exploring the practical benefits of argumentation in multi-agent deliberation
16 Koen Kok (VU), The PowerMatcher: Smart Coordination for the Smart Electricity Grid
17 Jeroen Janssens (UvT), Outlier Selection and One-Class Classification
18 Renze Steenhuizen (TUD), Coordinated Multi-Agent Planning and Scheduling
19 Katja Hofmann (UvA), Fast and Reliable Online Learning to Rank for Information Retrieval
20 Sander Wubben (UvT), Text-to-text generation by monolingual machine translation
21 Tom Claassen (RUN), Causal Discovery and Logic
22 Patricio de Alencar Silva (UvT), Value Activity Monitoring
23 Haitham Bou Ammar (UM), Automated Transfer in Reinforcement Learning
24 Agnieszka Anna Latoszek-Berendsen (UM), Intention-based Decision Support. A new way of representing and implementing clinical guidelines in a Decision Support System
25 Alireza Zarghami (UT), Architectural Support for Dynamic Homecare Service Provisioning
27 Mohammad Huq (UT), Inference-based Framework Managing Data Provenance
28 Frans van der Sluis (UT), When Complexity becomes Interesting: An Inquiry into the Information eXperience
29 Iwan de Kok (UT), Listening Heads
30 Joyce Nakatumba (TUE), Resource-Aware Business Process Management: Analysis and Support
31 Dinh Khoa Nguyen (UvT), Blueprint Model and Language for Engineering Cloud Applications
32 Kamakshi Rajagopal (OUN), Networking For Learning; The role of Networking in a Lifelong Learner’s Professional Development
33 Qi Gao (TUD), User Modeling and Personalization in the Microblogging Sphere
34 Kien Tjin-Kam-Jet (UT), Distributed Deep Web Search
35 Abdallah El Ali (UvA), Minimal Mobile Human Computer Interaction
36 Than Lam Hoang (TUE), Pattern Mining in Data Streams
37 Dirk Börner (OUN), Ambient Learning Displays
38 Elco den Heijer (VU), Autonomous Evolutionary Art
39 Joop de Jong (TUD), A Method for Enterprise Ontology based Design of Enterprise Information Systems
40 Pim Nijssen (UM), Monte-Carlo Tree Search for Multi-Player Games
41 Jochem Liem (UVA), Supporting the Conceptual Modelling of Dynamic Systems: A Knowledge Engineering Perspective on Qualitative Reasoning
42 Léon Planken (TUD), Algorithms for Simple Temporal Reasoning
43 Marc Bron (UVA), Exploration and Contextualization through Interaction and Concepts

2014 01 Nicola Barile (UU), Studies in Learning Monotone Models from Data
02 Fiona Tuliyano (RUN), Combining System Dynamics with a Domain Modeling Method
03 Sergio Raul Duarte Torres (UT), Information Retrieval for Children: Search Behavior and Solutions
04 Hanna Jochmann-Mannak (UT), Websites for children: search strategies and interface design - Three studies on children’s search performance and evaluation
05 Jurriaan van Reijsen (UU), Knowledge Perspectives on Advancing Dynamic Capability
06 Damian Tamburri (VU), Supporting Networked Software Development
07 Arya Adriansyah (TUE), Aligning Observed and Modeled Behavior
Chapter 12. SIKS dissertations

08 Samur Araujo (TUD), Data Integration over Distributed and Heterogeneous Data Endpoints
09 Philip Jackson (UvT), Toward Human-Level Artificial Intelligence: Representation and Computation of Meaning in Natural Language
10 Ivan Salvador Razo Zapata (VU), Service Value Networks
11 Janneke van der Zwaan (TUD), An Empathic Virtual Buddy for Social Support
12 Willem van Willigen (VU), Look Ma, No Hands: Aspects of Autonomous Vehicle Control
13 Arlette van Wissen (VU), Agent-Based Support for Behavior Change: Models and Applications in Health and Safety Domains
14 Yangyang Shi (TUD), Language Models With Meta-information
15 Natalya Mogles (VU), Agent-Based Analysis and Support of Human Functioning in Complex Socio-Technical Systems: Applications in Safety and Healthcare
16 Krystyna Milian (VU), Supporting trial recruitment and design by automatically interpreting eligibility criteria
17 Kathrin Dentler (VU), Computing healthcare quality indicators automatically: Secondary Use of Patient Data and Semantic Interoperability
18 Mattijs Ghijsen (UVA), Methods and Models for the Design and Study of Dynamic Agent Organizations
19 Vinicius Ramos (TUE), Adaptive Hypermedia Courses: Qualitative and Quantitative Evaluation and Tool Support
20 Mena Habib (UT), Named Entity Extraction and Disambiguation for Informal Text: The Missing Link
21 Kassidy Clark (TUD), Negotiation and Monitoring in Open Environments
22 Marieke Peeters (UU), Personalized Educational Games - Developing agent-supported scenario-based training
23 Eleftherios Sidirourgos (UvA/CWI), Space Efficient Indexes for the Big Data Era
24 Davide Ceolin (VU), Trusting Semi-structured Web Data
25 Martijn Lappenschaar (RUN), New network models for the analysis of disease interaction
26 Tim Baarslag (TUD), What to Bid and When to Stop
27 Rui Jorge Almeida (EUR), Conditional Density Models Integrating Fuzzy and Probabilistic Representations of Uncertainty
28 Anna Chmielowiec (VU), Decentralized k-Clique Matching
29 Jaap Kabbedijk (UU), Variability in Multi-Tenant Enterprise Software
30 Peter de Cock (UvT), Anticipating Criminal Behaviour
<table>
<thead>
<tr>
<th></th>
<th>Title</th>
<th>Author(s)</th>
</tr>
</thead>
<tbody>
<tr>
<td>31</td>
<td>Leo van Moergestel (UU), Agent Technology in Agile Multiparallel</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Manufacturing and Product Support</td>
<td></td>
</tr>
<tr>
<td>32</td>
<td>Naser Ayat (UvA), On Entity Resolution in Probabilistic Data</td>
<td></td>
</tr>
<tr>
<td>33</td>
<td>Tesfa Tegegne (RUN), Service Discovery in eHealth</td>
<td></td>
</tr>
<tr>
<td>34</td>
<td>Christina Manteli (VU), The Effect of Governance in Global Software</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Development: Analyzing Transactive Memory Systems.</td>
<td></td>
</tr>
<tr>
<td>35</td>
<td>Joost van Ooijen (UU), Cognitive Agents in Virtual Worlds: A</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Middleware Design Approach</td>
<td></td>
</tr>
<tr>
<td>36</td>
<td>Joos Buijs (TUE), Flexible Evolutionary Algorithms for Mining</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Structured Process Models</td>
<td></td>
</tr>
<tr>
<td>37</td>
<td>Maral Dadvar (UT), Experts and Machines United Against Cyberbullying</td>
<td></td>
</tr>
<tr>
<td>38</td>
<td>Danny Plass-Oude Bos (UT), Making brain-computer interfaces better:</td>
<td></td>
</tr>
<tr>
<td></td>
<td>improving usability through post-processing.</td>
<td></td>
</tr>
<tr>
<td>39</td>
<td>Jasmina Maric (UvT), Web Communities, Immigration, and Social</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Capital</td>
<td></td>
</tr>
<tr>
<td>40</td>
<td>Walter Omona (RUN), A Framework for Knowledge Management</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Using ICT in Higher Education</td>
<td></td>
</tr>
<tr>
<td>41</td>
<td>Frederic Hogenboom (EUR), Automated Detection of Financial Events</td>
<td></td>
</tr>
<tr>
<td></td>
<td>in News Text</td>
<td></td>
</tr>
<tr>
<td>42</td>
<td>Carsten Eijckhof (CWI/TUD), Contextual Multidimensional Relevance</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Models</td>
<td></td>
</tr>
<tr>
<td>43</td>
<td>Kevin Vlaanderen (UU), Supporting Process Improvement using</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Method Increments</td>
<td></td>
</tr>
<tr>
<td>44</td>
<td>Paulien Meesters (UvT), Intelligent Blauw. Met als ondertitel:</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Intelligence-gestuurde politiezorg in gebiedsgebonden eenheden.</td>
<td></td>
</tr>
<tr>
<td>45</td>
<td>Birgit Schmitz (OUN), Mobile Games for Learning: A Pattern-Based</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Approach</td>
<td></td>
</tr>
<tr>
<td>46</td>
<td>Ke Tao (TUD), Social Web Data Analytics: Relevance, Redundancy,</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Diversity</td>
<td></td>
</tr>
<tr>
<td>47</td>
<td>Shangsong Liang (UVA), Fusion and Diversification in Information</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Retrieval</td>
<td></td>
</tr>
</tbody>
</table>

2015 01 Niels Netten (UvA), Machine Learning for Relevance of Information in Crisis Response

02 Faiza Bukhsh (UvT), Smart auditing: Innovative Compliance Checking in Customs Controls

03 Twan van Laarhoven (RUN), Machine learning for network data

04 Howard Spoelstra (OUN), Collaborations in Open Learning Environments

05 Christoph Bösch (UT), Cryptographically Enforced Search Pattern Hiding
06 Farideh Heidari (TUD), Business Process Quality Computation - Computing Non-Functional Requirements to Improve Business Processes
07 Maria-Hendrike Peetz (UvA), Time-Aware Online Reputation Analysis
08 Jie Jiang (TUD), Organizational Compliance: An agent-based model for designing and evaluating organizational interactions
09 Randy Klaassen (UT), HCI Perspectives on Behavior Change Support Systems
10 Henry Hermans (OUN), OpenU: design of an integrated system to support lifelong learning
11 Yongming Luo (TUE), Designing algorithms for big graph datasets: A study of computing bisimulation and joins
12 Julie M. Birkholz (VU), Modi Operandi of Social Network Dynamics: The Effect of Context on Scientific Collaboration Networks
13 Giuseppe Procaccianti (VU), Energy-Efficient Software
14 Bart van Straalen (UT), A cognitive approach to modeling bad news conversations
15 Klaas Andries de Graaf (VU), Ontology-based Software Architecture Documentation
16 Changyun Wei (UT), Cognitive Coordination for Cooperative Multi-Robot Teamwork
17 André van Cleeff (UT), Physical and Digital Security Mechanisms: Properties, Combinations and Trade-offs
18 Holger Pirk (CWI), Waste Not, Want Not! - Managing Relational Data in Asymmetric Memories
19 Bernardo Tabuenca (OUN), Ubiquitous Technology for Lifelong Learners
20 Lois Vanhée (UU), Using Culture and Values to Support Flexible Coordination
21 Sibren Fetter (OUN), Using Peer-Support to Expand and Stabilize Online Learning
22 Zhemin Zhu (UT), Co-occurrence Rate Networks
23 Luit Gazendam (VU), Cataloguer Support in Cultural Heritage
24 Richard Berendsen (UVA), Finding People, Papers, and Posts: Vertical Search Algorithms and Evaluation
25 Steven Woudenberg (UU), Bayesian Tools for Early Disease Detection
26 Alexander Hogenboom (EUR), Sentiment Analysis of Text Guided by Semantics and Structure
27 Sándor Héman (CWI), Updating compressed column stores
28 Janet Bagorogoza (TiU), Knowledge Management and High Performance; The Uganda Financial Institutions Model for HPO
29 Hendrik Baier (UM), Monte-Carlo Tree Search Enhancements for One-Player and Two-Player Domains
30 Kiavash Bahreini (OU), Real-time Multimodal Emotion Recognition in E-Learning
31 Yakup Koç (TUD), On the robustness of Power Grids
32 Jerome Gard (UL), Corporate Venture Management in SMEs
33 Frederik Schadd (TUD), Ontology Mapping with Auxiliary Resources
34 Victor de Graaf (UT), Gesocial Recommender Systems

2016 01 Syed Saiden Abbas (RUN), Recognition of Shapes by Humans and Machines
02 Michiel Christiaan Meulendijk (UU), Optimizing medication reviews through decision support: prescribing a better pill to swallow
03 Maya Sappelli (RUN), Knowledge Work in Context: User Centered Knowledge Worker Support
04 Laurens Rietveld (VU), Publishing and Consuming Linked Data
05 Evgeny Sherkhonov (UVA), Expanded Acyclic Queries: Containment and an Application in Explaining Missing Answers
06 Michel Wilson (TUD), Robust scheduling in an uncertain environment
07 Jeroen de Man (VU), Measuring and modeling negative emotions for virtual training
08 Matje van de Camp (TiU), A Link to the Past: Constructing Historical Social Networks from Unstructured Data
09 Archana Nottamkandath (VU), Trusting Crowdsourced Information on Cultural Artefacts
10 George Karafotias (VUA), Parameter Control for Evolutionary Algorithms
11 Anne Schuth (UVA), Search Engines that Learn from Their Users
12 Max Knobbout (UU), Logics for Modelling and Verifying Normative Multi-Agent Systems
13 Nana Baah Gyan (VU), The Web, Speech Technologies and Rural Development in West Africa - An ICT4D Approach
14 Ravi Khadka (UU), Revisiting Legacy Software System Modernization
15 Steffen Michels (RUN), Hybrid Probabilistic Logics - Theoretical Aspects, Algorithms and Experiments
16 Guangliang Li (UVA), Socially Intelligent Autonomous Agents that Learn from Human Reward
17 Berend Weel (VU), Towards Embodied Evolution of Robot Organisms
18 Albert Meroño Peñuela (VU), Refining Statistical Data on the Web
19 Julia Efremova (Tu/e), Mining Social Structures from Genealogical Data
20 Daan Odijk (UVA), Context & Semantics in News & Web Search
21 Alejandro Moreno Célleri (UT), From Traditional to Interactive Playspaces: Automatic Analysis of Player Behavior in the Interactive Tag Playground
22 Grace Lewis (VU), Software Architecture Strategies for Cyber-Foraging Systems
23 Fei Cai (UVA), Query Auto Completion in Information Retrieval
24 Brend Wanders (UT), Repurposing and Probabilistic Integration of Data; An Iterative and data model independent approach
25 Julia Kiseleva (TU/e), Using Contextual Information to Understand Searching and Browsing Behavior
26 Dilhan Thilakarathne (VU), In or Out of Control: Exploring Computational Models to Study the Role of Human Awareness and Control in Behavioural Choices, with Applications in Aviation and Energy Management Domains
27 Wen Li (TUD), Understanding Geo-spatial Information on Social Media
28 Mingxin Zhang (TUD), Large-scale Agent-based Social Simulation - A study on epidemic prediction and control
29 Nicolas Höning (TUD), Peak reduction in decentralised electricity systems - Markets and prices for flexible planning
30 Ruud Mattheij (UvT), The Eyes Have It
31 Mohammad Khelghati (UT), Deep web content monitoring
32 Eelco Vriezekolk (UT), Assessing Telecommunication Service Availability Risks for Crisis Organisations
33 Peter Bloem (UVA), Single Sample Statistics, exercises in learning from just one example
34 Dennis Schunselaar (TUE), Configurable Process Trees: Elicitation, Analysis, and Enactment
35 Zhaochun Ren (UVA), Monitoring Social Media: Summarization, Classification and Recommendation
36 Daphne Karreman (UT), Beyond R2D2: The design of nonverbal interaction behavior optimized for robot-specific morphologies
37 Giovanni Sileno (UvA), Aligning Law and Action - a conceptual and computational inquiry
38 Andrea Minuto (UT), Materials that Matter - Smart Materials meet Art & Interaction Design
39 Merijn Bruijines (UT), Believable Suspect Agents; Response and Interpersonal Style Selection for an Artificial Suspect
40 Christian Detweiler (TUD), Accounting for Values in Design
41 Thomas King (TUD), Governing Governance: A Formal Framework
   for Analysing Institutional Design and Enactment Governance
42 Spyros Martzoukos (UVA), Combinatorial and Compositional Aspects
   of Bilingual Aligned Corpora
43 Saskia Koldijk (RUN), Context-Aware Support for Stress Self-
   Management: From Theory to Practice
44 Thibault Sellam (UVA), Automatic Assistants for Database Exploration
45 Bram van de Laar (UT), Experiencing Brain-Computer Interface Control
46 Jorge Gallego Perez (UT), Robots to Make you Happy
47 Christina Weber (UL), Real-time foresight - Preparedness for dynamic innovation networks
48 Tanja Buttler (TUD), Collecting Lessons Learned
49 Gleb Polevoy (TUD), Participation and Interaction in Projects. A Game-Theoretic Analysis
50 Yan Wang (UVT), The Bridge of Dreams: Towards a Method for Operational Performance Alignment in IT-enabled Service Supply Chains

2017 01 Jan-Jaap Oerlemans (UL), Investigating Cybercrime
02 Sjoerd Timmer (UU), Designing and Understanding Forensic Bayesian Networks using Argumentation
03 Daniël Harold Telgen (UU), Grid Manufacturing; A Cyber-Physical Approach with Autonomous Products and Reconfigurable Manufacturing Machines
04 Mrunal Gawade (CWI), Multi-core Parallelism in a Column-store
05 Mahdieh Shadi (UVA), Collaboration Behavior
06 Damir Vandic (EUR), Intelligent Information Systems for Web Product Search
07 Roel Bertens (UU), Insight in Information: from Abstract to Anomaly
08 Rob Konijn (VU) , Detecting Interesting Differences: Data Mining in Health Insurance Data using Outlier Detection and Subgroup Discovery
09 Dong Nguyen (UT), Text as Social and Cultural Data: A Computational Perspective on Variation in Text
10 Robby van Delden (UT), (Steering) Interactive Play Behavior
11 Florian Kunneman (RUN), Modelling patterns of time and emotion in Twitter #anticipointment
12 Sander Leemans (TUE), Robust Process Mining with Guarantees
13 Gijs Huisman (UT), Social Touch Technology - Extending the reach of social touch through haptic technology
14 Shoshannah Tekofsky (UvT), You Are Who You Play You Are: Modelling Player Traits from Video Game Behavior
15 Peter Berck (RUN), Memory-Based Text Correction
16 Aleksandr Chuklin (UVA), Understanding and Modeling Users of Modern Search Engines
17 Daniel Dimov (UL), Crowdsourced Online Dispute Resolution
18 Ridho Reinanda (UVA), Entity Associations for Search
19 Jeroen Vuurens (UT), Proximity of Terms, Texts and Semantic Vectors in Information Retrieval
20 Mohammadbashir Sedighi (TUD), Fostering Engagement in Knowledge Sharing: The Role of Perceived Benefits, Costs and Visibility
21 Jeroen Linssen (UT), Meta Matters in Interactive Storytelling and Serious Gaming (A Play on Worlds)
22 Sara Magliacane (VU), Logics for causal inference under uncertainty
23 David Graus (UVA), Entities of Interest — Discovery in Digital Traces
24 Chang Wang (TUD), Use of Affordances for Efficient Robot Learning
25 Veruska Zamborlini (VU), Knowledge Representation for Clinical Guidelines, with applications to Multimorbidity Analysis and Literature Search
26 Merel Jung (UT), Socially intelligent robots that understand and respond to human touch
27 Michiel Joosse (UT), Investigating Positioning and Gaze Behaviors of Social Robots: People’s Preferences, Perceptions and Behaviors
28 John Klein (VU), Architecture Practices for Complex Contexts
29 Adel Alhuraibi (UvT), From IT-Business Strategic Alignment to Performance: A Moderated Mediation Model of Social Innovation, and Enterprise Governance of IT"
30 Wilma Latuny (UvT), The Power of Facial Expressions
31 Ben Ruijl (UL), Advances in computational methods for QFT calculations
32 Thaer Samar (RUN), Access to and Retrievability of Content in Web Archives
33 Brigit van Loggem (OU), Towards a Design Rationale for Software Documentation: A Model of Computer-Mediated Activity
34 Maren Scheffel (OU), The Evaluation Framework for Learning Analytics
35 Martine de Vos (VU), Interpreting natural science spreadsheets
36 Yuanhao Guo (UL), Shape Analysis for Phenotype Characterisation from High-throughput Imaging
37 Alejandro Montes Garcia (TUE), WiBAF: A Within Browser Adaptation Framework that Enables Control over Privacy
38 Alex Kayal (TUD) Normative Social Applications
39  Sara Ahmadi (RUN) Exploiting properties of the human auditory system and compressive sensing methods to increase noise robustness in ASR
40  Altaf Hussain Abro (VUA) Steer your Mind: Computational Exploration of Human Control in Relation to Emotions, Desires and Social Support For applications in human-aware support systems
41  Adnan Manzoor (VUA) Minding a Healthy Lifestyle: An Exploration of Mental Processes and a Smart Environment to Provide Support for a Healthy Lifestyle


M. G. Kendall. Rank correlation methods. 1948.


