RADBOUD UNIVERSITY NIJMEGEN

Combining causal models for biomedical applications

Master thesis

Author:	Lisanne Kluitmans
Student number:	S4259386
Study:	Information Science
Supervisor 1:	Tom Heskes
Supervisor 2:	Patrick van Bommel

Nijmegen, April 2015

CONTENTS

1. MOTIVATION

This master thesis is written at the Radboud University in Nijmegen for completion of the master study Information Science. The subject for this master thesis is combining causal models for biomedical applications and the existing methods to combine causal models from several different experiments and their differences.

I chose to do my master thesis about combining causal models for biomedical applications because of my previous study Bio-Informatics at the HAN in Nijmegen. I have always been interested in the combination of biology with informatics because of the many useful applications that can be created. At the moment there are, for example, very large amounts of medical data available about patients, so it is important to be able to handle this large amount of data correctly and efficiently and to be able to draw the right conclusions from it. Though, after graduating from Bio-Informatics, I chose to do a master study that was more on the informatics side (Information Science) rather than the biology side. I made this decision because I wanted to learn more about this interesting field and to have more experience with informatics, but I never lost interest in the combination of informatics with biology.

When I had to choose a subject for my master thesis, I knew that I wanted to find a subject that also has a bit of a biology side to it. I started to talk about this with Tom Heskes and he proposed this topic about combining causal models for biomedical applications and I immediately was very interested to work on this.

2. INTRODUCTION

This master thesis focuses on combining causal models for biomedical applications and the existing methods to combine causal models from different experiments and their differences. This master thesis will explain what causal models exactly are, what different kind of causal models there are at the moment and how these causal models can be used in biomedical applications. Also, the pitfalls in using causal models will be discussed and the advantages in using causal models will be explained.

There are a lot of different methods available to create causal models from the variables in a single dataset (Tillman R. E., 2014). But in practice it often occurs that you want to combine the variables of two or more datasets, including datasets that do not measure all the variables of interest, but have an overlap with other datasets. This is not possible with the methods that are used to work with a single dataset. At the moment, it is very difficult to find or create a good working and reliable method to combine causal models from several different experiments to get a clear and more informative overview of the results. So, the goal of this master thesis is to find out what different kind of methods there are to combine causal models from several different experiments and what their differences are. In this way there will be a clear overview of those different methods and their differences.

The third chapter will explain more about what a causal model exactly is. The fourth chapter will explain what kind of different causal models there are at the moment. Then, the advantages and the pitfalls of using causal models will be discussed in the fifth and the sixth chapter. After that, the seventh chapter will explain how causal models can be used in biomedical applications. The eighth chapter will present the findings about the different methods to combine causal models from different experiments. The ninth chapter contains a case study. The tenth chapter will present the differences between the methods. Finally, in the last chapter, there will be the conclusion of this master thesis.

Four sub questions have been formulated about causal models that will be answered in this master thesis. If these questions are answered, there will be a clear overview of the aspects of causal models at the moment. Also, I have formulated three main questions, which will be answered in the eighth and tenth chapter. To get an answer on these main questions is the main goal of this master thesis.

2.1. SUB QUESTIONS

- What kind of causal models are there?
- What are the advantages of causal models?
- What are the pitfalls of causal models?
- How are causal models applied in biomedical applications?

2.2. MAIN QUESTIONS

- What methods are available to combine causal models from different experiments?
- What are the differences of the methods to combine causal models from different experiments?
- Which method is to be preferred in which situations?

3. CAUSAL MODELS

Causal models are used in a lot of different scientific domains, for example in information technology (Grant, 1991), psychology, economics, social sciences (Steel, 2011) and (Russo, 2011), and in biology (Le, 2004) and (Kleinberg, 2011). Causal models can also be used for the decision making processes in, for example, important business decisions in a company: causal models can clearly show what the effects of implementing a new policy are. This master thesis focuses solely on causal models in bio-medical applications.

A causal model shows which variable is the cause of another variable and what the effect of that variable is on another variable, that it is why these models are called causal models. In this way the causal relationships between the different variables of interest can be seen easily and clearly. With the help of causal models you can also see what happens to the other variables of interest if one of the variables would be missing or would be knocked out and predictions can be made about the effect of such an event on the other variables in the causal model.

3.1. WHAT IS A CAUSAL MODEL?

A causal model is a graph in which the causal relationships between the different variables of interest are shown, so you can see which of the variables have an influence on which other variables. A causal model typically consists of nodes, representing the variables, and directed edges, indicating the interactions (causal relationships) between those different variables. Figure 1 shows an example of a causal model. Causal models can become very large and very complex with a lot of different variables and relationships. (Figure 1 is a very simple and small example of a causal model.)

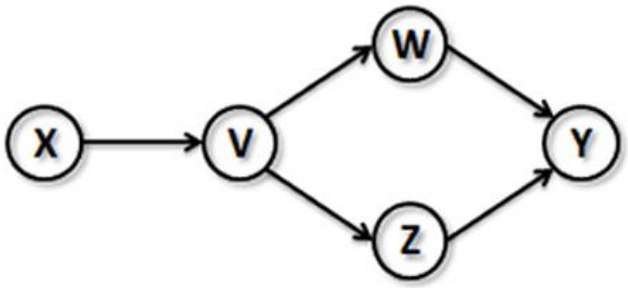

Figure 1 – Example of a causal model

In figure 1 nodes *V*, *W*, *X*, *Y* and *Z* are the different variables of this causal model. The arrowheads on the edges show the directions of the causal relationships between these variables. Two variables have a causal relationship if there is a directed path from one of these variables to the other variable. If this is not the case, these two variables have a noncausal relationship (T. Claassen, 2010).

These are the causal relationships in the causal model in figure 1:
- Variable *X* has a causal effect on variable *V*, so variable *X* is the parent of variable *V* and that makes variable *V* a child of variable *X*.
- Variable *V* has a causal effect on both variable *W* and variable *Z*, so variable *V* is the parent of both

variable *W* and variable *Z* and that makes variable *W* and variable *Z* children of variable *V*.
- Variable *W* and variable *Z* both have a causal effect on variable *Y*, so variable *Y* has two parents: variable *W* and variable *Z* and this makes variable *Y* a child of both variable *W* and variable *Z*.

Variables can also have ancestors and descendants (T. Claassen, 2010). In figure 1, variable *X* is an ancestor of variable *W*, variable *Z* and variable *Y*. Variable *W*, variable *Z* and variable *Y* are descendants of variable *X*. Variable *V* is also an ancestor of variable *Y* and variable *Y* is thus a descendant of variable *V*. So, a variable can have parents, children, ancestors and descendants, but variables can also have non-descendants. The non-descendants of a variable are the variables in the causal model that are not a descendant of that particular variable (Daly, 2011).

There are direct causal relationships between the variables and indirect causal relationships between the variables in a causal model. An example of a direct causal relationship is the relationship between variable *Z* and variable *Y* in figure 1. A direct causal relationship means that there is no intervening variable between variable *Z* and variable *Y*: variable *Z* has a direct causal effect on variable *Y*. These two variables are also adjacent, this means that there is an edge between those two variables (T. Claassen, 2010). An example of an indirect causal relationship is the relationship between variable *V* and variable *Y* in figure 1. An indirect causal relationship means that variable *V* does have a causal influence on variable *Y*, but this influence goes through another variable, in this case through variable *Z*. Variable *V* has an indirect causal effect on variable *Y*.

Causal models can be used to reason about the effect of an intervention or a manipulation on the other variables in the causal model. In biology, causal models are used to show, for example, which genes have an effect on which other genes, or what the effect is on the other genes if a particular gene would be knocked out. This can help in finding the cause of a disease and in the development of a medicine for this disease (Kleinberg, 2011). Chapter 7 about causal models in biomedical applications will explain more about causal models in biomedical applications.

3.2. CONDITIONAL (IN)DEPENDENCE

Two variables in a causal model can be conditionally independent or conditionally dependent given a third variable. These three variables together are called a triple. This also holds for sets of variables. If the conditional independence information about the variables of interest is known, a causal model can be constructed from this information. There are three kinds of conditional independencies: tail-tail, head-tail, and head-head. Figure 2 is an example of a tail-tail conditional independence. Variable *C* stands for cold, variable *S* stands for sneezing and variable *R* stands for runny nose. This causal model shows that a cold can cause someone to sneeze and to have a runny nose. If you want to know if sneezing and having a runny nose are conditionally independent or not, you have to condition on variable *C*. This means that if you know that someone has a cold, you know that this person can sneeze and can have a runny nose. These two variables depend on having a cold, and not on each other, this is why these two variables are conditionally independent conditional on a cold. This is called a tail-tail conditional independence because of the orientation of the edges towards the variable that is conditioned on. In this case, the tails of the two edges are both at the variable that is conditioned on.

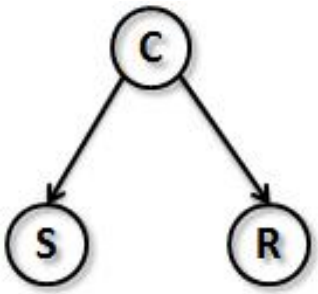

Figure 2 - Tail-tail: variable *S* and variable *R* are conditionally independent given variable *C*

Figure 3 is an example of a head-tail conditional independence. Variable *S* stands for smoking, variable *L* stands for lung cancer and variable *X* stands for a positive X-ray. This causal model shows that smoking can cause lung cancer and lung cancer can cause a positive X-ray. If you want to know if smoking and a positive X-ray are conditionally independent or not, you have to condition on variable *L*. This means that if you know that someone has lung cancer, you know that this person can have a positive X-ray, but knowing if this person smokes or not will not affect the positive X-ray anymore. Therefore, smoking and a positive X-ray are conditionally independent conditional on lung cancer. In this case, one of the two edges has an arrowhead at the variable that is conditioned on and the other edge has a tail at the variable that is conditioned on. This is why this is called a head-tail conditional independence.

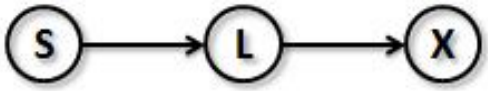

Figure 3 - Head-tail: variable *S* and variable *X* are conditionally independent given variable *L*

Figure 4 is an example of a head-head conditional independence. Variable *F* stands for the flu, variable *A* stands for an allergy and variable *S* stands for sinus inflammation. This causal model shows that having the flu and having an allergy can cause a sinus inflammation. If you want to know if having the flu and having an allergy are conditionally independent or not, you have to condition on variable *S*. This means that if you know that someone has a sinus inflammation, you know that this can come from having the flu or having an allergy. So, variable *F* and variable *A* become dependent. In this case, the arrowheads of the two edges are both at the variable that is conditioned on. This is why this is called a head-head conditional independence.

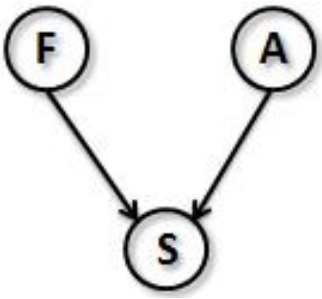

Figure 4 - Head-head: variable *F* and variable *A* are conditionally dependent given variable *S*.

3.3. D-SEPARATION

The ‘d’ in d-separation stands for directional (Geiger, 1990), because it is applied in directed causal models. With d-separation you can directly read the conditional (in)dependencies from the causal model. If variable *X* and variable *Y* are d-separated given a variable *Z*, variable *X* and variable *Y* are also conditionally independent given variable *Z*. Also, Variable *X* and variable *Y* are d-separated given variable *Z* if there are no active paths between variable *X* and variable *Y* (Geiger, 1990). A path can be divided into triples. Figure 5 shows when a triple is active or inactive. If there is one inactive triple in a path, then the whole path is inactive. A path becomes active if each triple in that path is active. The shaded variables in figure 5 mean that these are variables that are conditioned on.

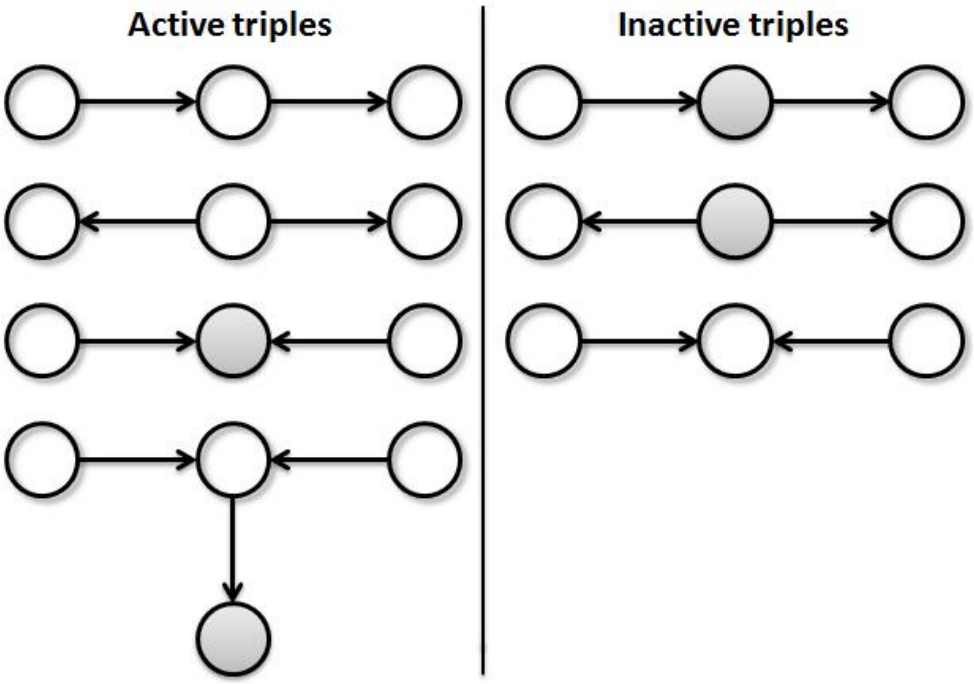

Figure 5 - Active and inactive triples (the shaded variables are variables that are conditioned on)

Variable *X* and variable *Y* are also d-separated given variable *Z* if all paths from variable *A* to variable *B* are blocked by variable *Z*. If variable *X* and variable *Y* are not d-separated, they are d-connected. This is the opposite of d-separation. These conditions do not only hold for single variables, but also for sets of variables. For MAGs m-separation in used, this is the same as d-separation, but then applied to MAGs (Richardson, 2002).

4. TYPES OF CAUSAL MODELS

There are different kinds of causal models that represent different kind of causal relationships, but all those causal models look quite similar. This chapter will explain the differences between those causal models and their causal relationships.

These are the different types of causal models that will be discussed here:
- DAG (Directed Acyclic Graph)
- MAG (Maximal Ancestral Graph)
- CPAG (Complete Partial Ancestral Graph)

4.1. DAG

A Directed Acyclic Graph (DAG) is a graph that does not contain a cycle, this is called acyclic. This means that there is no path from one particular variable *X* to the same variable *X*. A DAG is also directed, this means that the edges all have an arrowhead to show which variable has an influence on which other variable. These edges are called directed edges. Figure 1 from the previous chapter is also an example of a DAG. A DAG represents both the observed variables and the unobserved, hidden variables and the relationships between these different kinds of variables. Observed variables are variables that can be measured, like for example gender and age. Unobserved, hidden variables are variables that cannot be (easily) measured, like for example behavior and happiness. Although the unobserved, hidden variables cannot be measured, they can have a big influence on the other variables in a causal model. So, it is very important to know if there are unobserved, hidden variables or to allow for the possibility of there being unobserved, hidden variables and to include them in the causal model. There is a difference between a DAG and a causal DAG. If a DAG is called a causal DAG, it can be used to compute the effect of interventions and of manipulations on the other variables of interest. A DAG shows if there is a relationship between the different variables, but a causal DAG shows the causal relationships between the different variables of interest in the causal model.

A DAG is the graph structure of a Bayesian network. A Bayesian network is a DAG with probability distributions in tables called Conditional Probability Tables (CPT's). A Bayesian network looks the same as a DAG, it also contains variables and edges and a Bayesian network is also acyclic and directed, but the meaning of a Bayesian network is a little bit different than the meaning of a DAG. In a Bayesian network the probability of one variable *X* is shown in the CPT's given another variable *Y*. So, a Bayesian network does not only consist of variables and edges, but every variable in the graph has its own probability distribution (Steel, 2011). A DAG does not show the probability with which an effect occurs, but a Bayesian network does show this. Figure 6 shows an example of a Bayesian network (Le, 2004). This figure shows that gene A and gene B both have a probability of 0.5 of being on (1). The combination of gene A and gene B has a positive effect on gene C. The table of gene C shows what the probabilities of gene C are in all the possible combinations of gene A and gene B (on or off). Gene C has a negative effect on gene D. The table of gene D shows that when gene C is off (0), the probability of gene D being on is 0.9. When gene C is on, the probability of gene D being off is 0.

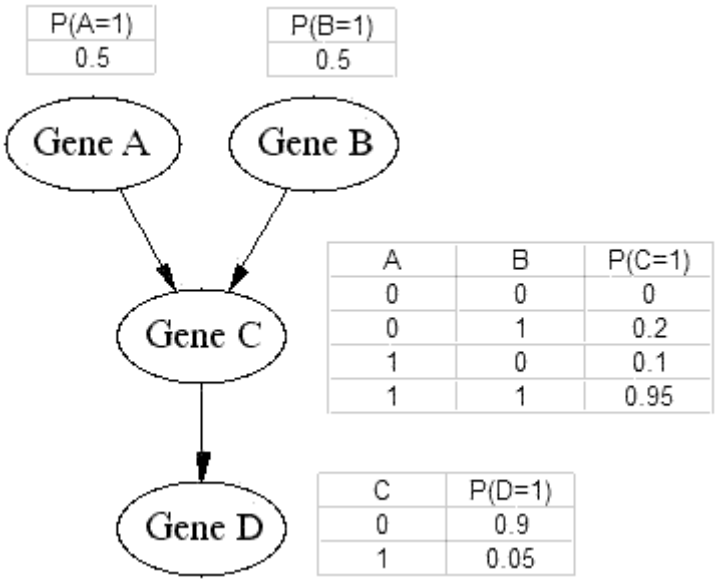

Figure 6 – Example of a Bayesian network

4.2. MAG

A Maximal Ancestral Graph (MAG) also looks quite similar to a DAG and a MAG is also acyclic. A MAG only represents the observed variables, not the unobserved, hidden variables (Tillman R. E., 2014). Out of every DAG a MAG can be created if it is known which variables are the observed variables and which variables are the unobserved, hidden variables. A MAG contains two extra types of edges besides the edges in the form of arrowheads (Tillman R. E., 2011). A MAG also contains edges with arrowheads on both ends, these edges are called bi-directed edges. A MAG also contains edges with no arrowheads at all, these edges are called undirected edges. A bi-directed edge means that there is a hidden common cause between those two variables. This means that there is one other variable that causes these two variables. An undirected edge means that you know that there is a causal relationship between those two variables, but you do not know which variable has an effect on which variable.

4.3. CPAG

A Complete Partial Ancestral Graph (CPAG) looks similar to a MAG, but a CPAG contains one other extra type of edge. A CPAG contains edges with a circle mark at the end (Zhang, 2008). This circle mark represents a combination of an undirected edge and an arrowhead. In figure 8 (Claassen, 2010) there are four examples of different CPAGs. A CPAG also only represents the observed variables and not the unobserved, hidden variables. A CPAG can be created directly from the dataset, this is why CPAGs are used as an input for methods to combine causal models from different experiments. From one single CPAG several different MAGs can be created. This is the case because of the circle mark edges: you do not know for sure if variable *X* has an influence on variable *Y* or the other way around. So, the several different MAGs show the different options of possible edges.

5. ADVANTAGES OF CAUSAL MODELS

Besides the few disadvantages, causal models also have important advantages. These advantages outweigh the disadvantages of working with causal models. This chapter will explain the great advantages in using causal models.

These are the advantages in working with causal models that will be discussed here:
- Easy to understand
- Predictions and interventions
- Prior knowledge
- Visualization

5.1. EASY TO UNDERSTAND

Causal models are very easy to understand for everyone, even if the causal model becomes very large and complex. A causal model is way easier to understand than a large piece of complicated text that explains the same causal relationships, because you only need to know the meaning of the (different) types of edges in a causal model and then you immediately see all the causal relationships between the different variables of interest.

5.2. PREDICTIONS AND INTERVENTIONS

Causal models can be used to do predictions about certain events (Sillignakis, 2001), (Russo, 2011), and (Kleinberg, 2011). If you know the value of a particular variable *X*, you can predict the effect of this variable *X* on the other variables of interest in the causal model. If, for example, gene A from figure 6 has a mutation that causes this gene to be off all the time, this would have an important effect on gene C and gene D. This would mean that gene C only has a probability of 0.2 to be on if gene B is also on. Because of this low probability of gene C being on, gene D has an even higher probability of being on. Gene B will not be affected by this mutation in gene A. Causal models are also used to do predictions about for example the weather and oil prices (Daly, 2011).

5.3. PRIOR KNOWLEDGE

It is possible to include prior knowledge about causal relationships into a causal model (Borboudakis, 2012). This prior knowledge can, for example, be about the absence of particular causal relationships between the variables of interest or about the presence of particular causal relationships in the causal model. This makes it easier and faster to create the causal model.

5.4. VISUALIZATION

Causal models are also very clearly visualized. They only contain the variables of interest (nodes) and the causal relationships between those variables of interest (edges). There are no large pieces of text or complicated formulas that you have to try to understand.

6. PITFALLS OF CAUSAL MODELS

Causal models are a great way to clearly visualize data and to show the causal relationships between the different variables of interest, but using causal models also has a few pitfalls (Magidson, 1982). This chapter will explain those pitfalls in working with causal models.

These are the disadvantages in working with causal models that will be discussed here:
- Unobserved, hidden variables
- Omitted relations between variables
- Complexity
- Misinterpretation
- Missing data

6.1. UNOBSERVED, HIDDEN VARIABLES

It is possible that an unknown variable *X* is influencing other variables in the dataset while you do not know that this variable *X* can have an effect on these other variables. This variable *X* is called an unobserved, hidden variable. If these unobserved, hidden variables are not shown in the causal model, the causal model is not complete and the conclusions that are drawn from this causal model can be incorrect or incomplete. Two variables can have a hidden common cause, this means that there is a unobserved, hidden variable that has a causal effect on those two variables. This hidden common cause can, for example, be a variable that you did not know that could have an effect on the other variables, or it can be a variable that you forgot to take into account for this causal model. Figure 7 shows an example of a situation with a hidden common cause (Neapolitan, 2007). This causal model shows that if someone has a cold (*C*), this causes them to sneeze (*S*) and to have a runny nose (*R*). So, if someone is sneezing and/or has a runny nose, you would say that this person has a cold. But a cold is not the only variable that can cause someone to sneeze and to have a runny nose. There is a hidden common cause that can cause the same symptoms as a cold does. This hidden common cause is hay fever (*H*). Without this hidden common cause the wrong conclusion would be drawn from this causal model: only a cold can cause someone to sneeze or to have a runny nose. This is not the right conclusion because hay fever can also cause someone to sneeze or to have a runny nose. This is why it is important to take the unobserved, hidden variables into account in a causal model.

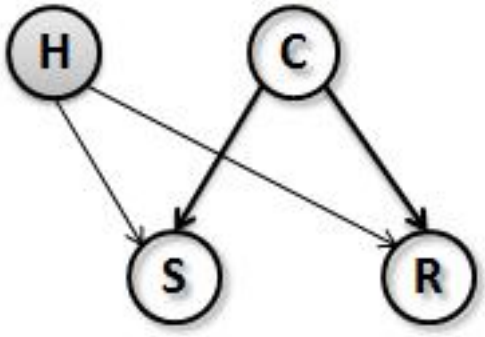

Figure 7 – Example of a hidden common cause (variable H)

It can also be the case that because of an omitted variable *X* the causal relationships between some other variables in the causal model are not correct and the wrong conclusions are drawn from the causal model. So, it is very important to know if there are unobserved, hidden variables or to allow for the possibility of there being unobserved, hidden variables and to include them in the causal model. It is hard to decide which variables to include in the dataset and which variables not need to be included in the dataset, especially when the dataset is already very large. It can happen that you forget to include one important variable or that you did not know that a particular variable should have been included in the dataset. In this way you also miss a variable in the causal model which can have large consequences for the causal relationships in the causal model and the conclusions that will be drawn from it.

6.2. OMITTED RELATIONS BETWEEN VARIABLES

Sometimes a causal relationship between two variables can be omitted by accident. This can also have a large effect on the other relationships in the causal model. This can for example also happen because of an unobserved, hidden variable. Then, again, the causal model will be incorrect and the wrong conclusions can be drawn from this causal model.

6.3. COMPLEXITY

Causal models can contain a lot of different variables and relationships and because of this they can become very large and complex. This can be the case in, for example, biology, when you are modeling the effects of a disease in which a large number of genes are involved. If a causal model consists of a lot of different variables of interest and becomes very large, it can become very confusing and difficult to clearly see the causal relationships between all the different variables of interest.

6.4. MISINTERPRETATION

A causal model can be misunderstood. If someone does not know the precise meaning of the different kind of edges (directed, bi-directed, undirected and circle mark), this person can conclude different things from the causal model than he should conclude. Causal models can also have different meanings (Daly, 2011). Also, noncausal DAG's can be mistaken for causal DAG's. So, it is important to know what the background information of the causal model is to correctly understand the meaning of the causal model and the relationships between the different variables of interest and to avoid misinterpretation.

6.5. MISSING DATA

Most datasets contain missing values (Daly, 2011). Missing values are values that are not known, but these unknown values can have a large effect on the relationships in the causal model. Dealing with these missing values is very difficult. There are a few methods on how to work with missing data, but missing data stays a big problem in working with causal models.

7. CAUSAL MODELS IN BIOMEDICAL APPLICATIONS

Causal models are applied in a lot of different scientific domains, as mentioned before, because they are very easy to use and clear to understand. Also in biomedical applications causal models are a great help.

7.1. IMPORTANCE

In biomedical applications causal models can be used to show, for example, which genes have an effect on which other genes, as shown in the example model in figure 6. This can help in finding the cause of a disease and in the development of a medicine for this disease (Kleinberg, 2011). Causal models can also help in finding the cause of some particular symptoms, as shown in the example model in figure 7, or causal models can show what the effects of the use of a particular medicine are, or what the effect is on the other genes in the causal model if, for example, one gene would be knocked out. In this way, predictions about certain events can be made.

7.2. BIOLOGY

In biology, there are a lot of important processes in which causal models are being used (Kleinberg, 2011). Also, being able to combine the information from several different datasets is very important in biology. This is often the case in, for example, functional magnetic resonance imaging (fMRI) studies (Ramsey, 2010). fMRI studies are used, for example, to identify the causal relationships between different regions of the brain. These studies are often performed on multiple subjects. In this way there will be multiple datasets (one for each subject) that have to be combined to create one causal model to clearly see the causal relationships between the different regions of interest (ROI) in the brain.

7.3. EXAMPLE

In chapter 9 the procedure, results, and pitfalls of a case study I did are shown. This case study is an example of how causal models can be used in relations to biology. The case study is about combining the causal models from three different datasets containing data about ADHD (attention deficit hyperactivity disorder) patients.

8. METHODS TO COMBINE CAUSAL MODELS FROM DIFFERENT EXPERIMENTS

There are several different methods available at the moment that can combine causal models from different experiments. It is very useful to use such a method because then there will be one clear causal model that contains all the causal information about the variables of interest from all the different experiments, instead of multiple different causal models that have to be compared to each other. In this chapter the different methods to combine causal models from several different experiments will be explained.

These are the methods for combining causal models that will be discussed here:
- MCI-algorithm
- ION algorithm
- IOD algorithm
- cSAT+ algorithm

Figure 8 (Claassen, 2010) explains what a method for combing causal models from several different experiments can do. A number of different causal models is used as an input (models 1, 2, and 3) and one causal model that contains the causal information of all the different variables of interest from the input models combined is the output of the method (the last causal model in figure 8).

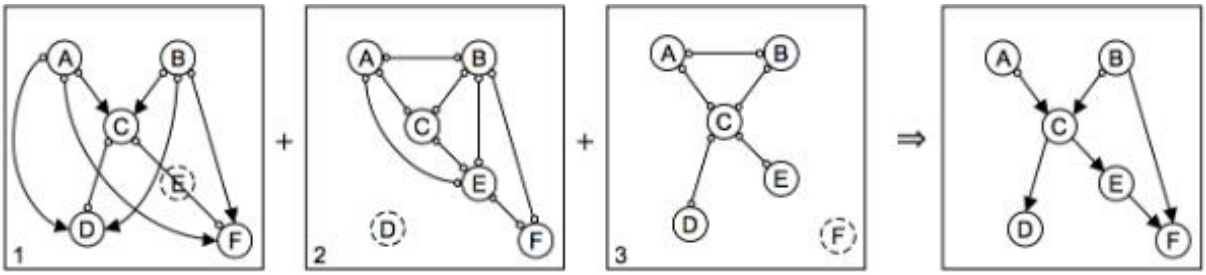

Figure 8 - Three different causal models combined to one, more informative, causal model

8.3. MCI-ALGORITHM

The Multiple model Causal Inference (MCI) algorithm (Claassen, 2010) is a reliable and fast method to combine the causal information from several different experiments. The MCI-algorithm works with a set of CPAGs from several different experiments as an input and gives an easily interpretable and a very clear output. The output of the MCI-algorithm is in the form of one causal model, which is very concise. Suppose you have two different datasets which contain the different variables of interest. Dataset one contains variable *A*, variable *B*, variable *C*, variable *E* and variable *F*. Dataset two contains variable *A*, variable *B*, variable *D* variable *E* and variable *F*. So, in this case, the input for the MCI-algorithm is in the form of two different CPAGs, one for every dataset. First, the MCI-algorithm creates a graph G containing all the variables of interest from the different datasets. In this graph G, all the variables are connected with edges that have circle marks on both ends. This is called a fully con-

nected graph. Figure 9 (1, 2, and 3) is an example of this first step of the MCI-algorithm. The first two causal models in figure 12 represent the two input CPAGs from the two datasets and the third causal model shows the fully connected graph of the variables of the two input CPAGs combined. The dashed circles represent the variables that are not present in that dataset.

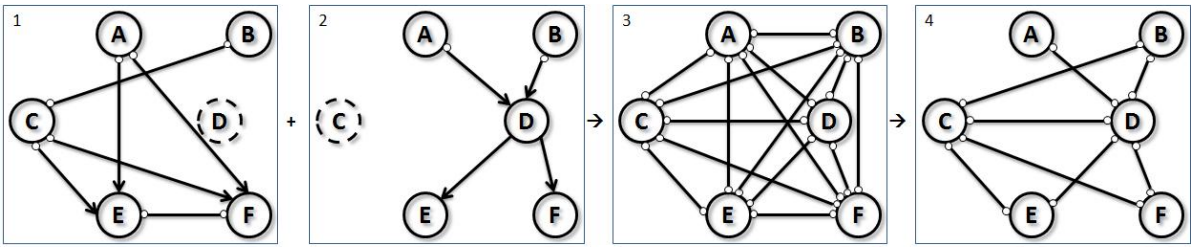

Figure 9 - 2 input CPAGs (1 and 2) are combined to one fully connected graph (3) and after that the edges that do not appear between the variables of interest are removed (4) (MCI-algorithm)

Then, all the edges that do not appear between the variables of interest in the input CPAGs are removed from the graph G. For example, there is no edge between variable *A* and variable *B* in both of the input CPAGs, so this edge is removed from the graph G. The last causal model in figure 9 shows what graph G looks like after this step. In the next step, a causal relations matrix M_c is created. This matrix will show all the (non)causal relationships between the different variables of interest. If there is a causal relationship from one variable to another variable in the input CPAGs, then there will be a green box in the M_c matrix for that relationship between those two variables. If there is no causal relationship from one variable to another variable in one of the input CPAGs, then there will be a red box in the M_c matrix for that relationship between those two variables. If you do not know yet what kind of causal relationship there is between two variables (a circle mark), then there will be a gray box in the M_c matrix for that relationship between those two variables. In this way, prior knowledge about the causal relationships between the different variables of interest can be shown in the M_c matrix. Figure 10 shows what the M_c matrix looks like for the causal relationships between the variables of interest in the example of figure 9.

	A	B	C	D	E	F
A						
B						
C						
D						
E						
F						

Figure 10 – M_c matrix, the green boxes represent a causal relationship, the red boxes show that there is no causal relationship between those variables, and the gray boxes show that there can be a causal relationship (MCI-algorithm)

After this, another matrix (a combined SCI-matrix) is used to represent the causal information about

all the triples, this is the most expensive step of the MCI-algorithm. Figure 11 represents an example of a combined SCI-matrix.

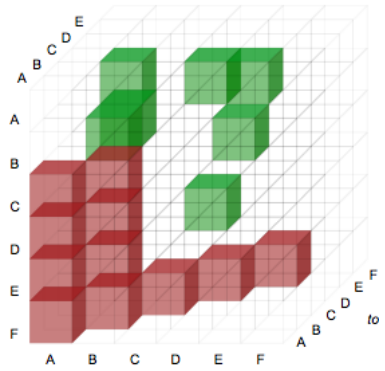

Figure 11 – Example of a combined SCI-matrix, representing causal information about triples (MCI-algorithm)

Finally, in the last step, the MCI-algorithm combines the causal information from the M_c matrix and the SCI-matrix and translates this causal information to edges and endpoint orientations in the graph G. This graph is the output causal model of the MCI-algorithm. The MCI-algorithm works very well in the large sample limit and works for different sets of causal models that consist of up to 20 nodes (Claassen, 2010).

8.4. ION ALGORITHM

The Integration of Overlapping Networks (ION) algorithm (R. Tillman, 2009) is a complete and sound method that can combine causal models from several different experiments. To the best of my knowledge, the ION algorithm is the first causal algorithm created that can work with variables from several different datasets. The ION algorithm also uses a set of CPAGs as input, one for each dataset. The output of the ION algorithm is in the form of another set of CPAGs. First, the ION algorithm creates a fully connected graph K that contains all the variables of interest from the different datasets. This first step is the same as the first step from the MCI-algorithm (see figure 9). Then, all the nonadjacencies and endpoint orientations of the edges in the input CPAGs are transferred to the graph K. So, for example, the edge between variable *X* and variable *Y* in graph K will be removed if variable *X* and variable *Y* are not adjacent in the input CPAGs. This is what also happened to the edge between variable *A* and variable *B* in the example from figure 9. Also, the endpoint at variable *Y* from the edge between variable *X* and variable *Y* will become an arrowhead in graph K if variable *X* causes variable *Y* in the input CPAGs. Figure 12 shows what the graph K looks like after this step with the same variables of interest as from the example in figure 9.

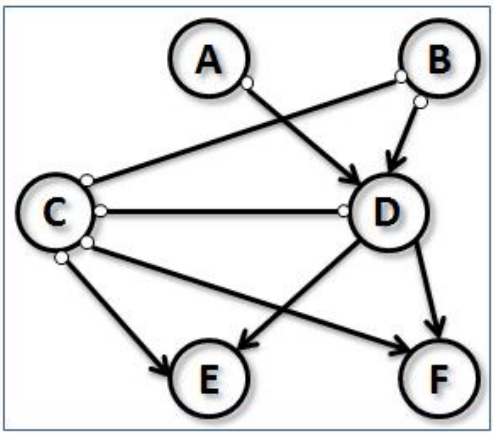

Figure 12 - Fully connected graph with endpoint orientations and removed edges (ION algorithm)

After that, all possibly active paths between all the possible pairs of variables are recorded in a table called PAT (Possibly Active Trails). If you change the endpoint orientations of one or more edges in a path between two variables from a circle mark to a tail or an arrowhead and the path becomes active because of this change, this is called a possibly active path. Then, all minimal hitting sets are recorded in another table called PC. Minimal hitting sets are changes that can be made to a path that makes this path no longer active. The next step is to create a graph A_i of each minimal hitting set. If this graph A_i is consistent with every input CPAG, this graph A_i is added to the set of output CPAGs A. If all minimal hitting sets are checked and added to the set of output CPAGs A, this set A is the output of the ION algorithm. The ION algorithm is a complete and sound algorithm, but it does not work well with many variables of interest (Tillman R. E., 2014) and it performs a lot of (brute force) operations. Also, the ION algorithm is a computationally hard algorithm, even when there is a small number of different variables of interest. The more variables of interest there are, the more likely it is that there will be more CPAGs in the output CPAG set. In this way the output of the ION algorithm can become very large and this makes it hard to deal with the output.

8.5. IOD ALGORITHM

The Integration of Overlapping Datasets (IOD) algorithm (Tillman R. E., 2011) is another correct and complete algorithm that can combine the variables of interest from several different datasets. The IOD algorithm works with a set of different datasets as input and also gives a set of CPAGs as output. The IOD algorithm is a complex version of the ION algorithm. The IOD algorithm is more accurate and robust than the ION algorithm and the IOD algorithm outperforms the ION algorithm in precision and recall. The IOD algorithm works in a different way as the MCI-algorithm and the ION algorithm. The IOD algorithm deals with all the data at once instead of dealing with several different input CPAGs. The IOD algorithm combines all the variables of interest and tests for conditional independence between all the possible pairs of variables. Each conditional independence test gets a p-value assigned to it. So, for example, dataset one and dataset two both contain variable *X* and variable *Y*. First, variable *X* and variable *Y* are tested on conditional independence in dataset one, then variable *X* and variable *Y* are tested on conditional independence for dataset two. The two resulting p-values are combined and one p-value is calculated for these two variables. Based on this combined p-value, a

conclusion is drawn if variable *X* and variable *Y* are independent or not. In this way the IOD algorithm can handle contradictory information in the datasets. The IOD algorithm requires less memory and is much faster than the ION algorithm (Tillman R. E., 2014).

8.6. cSAT+ ALGORITHM

The causal SAT+ algorithm (Triantafilou, 2010) is a different method to combine causal models from several different variable datasets. The cSAT+ algorithm is much faster than the ION algorithm. The cSAT+ algorithm uses a set of CPAGs as input and returns a Pairwise Causal Graph (PCG) as output. A PCG is a type of graph that contains two different types of edges (Tsamardinos, 2012): dashed edges and solid edges. The PCG contains, just like the CPAG, three types of endpoint orientations to these edges: the arrowhead, the tail, and the circle mark. In figure 13 an example of a PCG is shown (Tsamardinos, 2012).

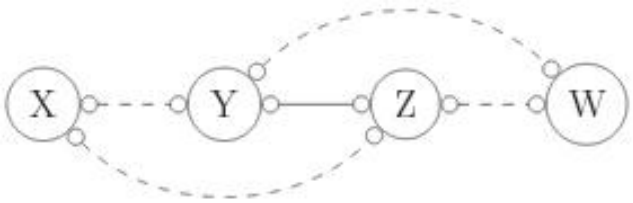

Figure 13 – Example of a Pairwise Causal Graph (cSAT+ algorithm)

First, the cSAT+ algorithm generates a complete unoriented graph U containing all the variables of interest from the input CPAGs. Then, all the missing edges and all the endpoint orientations of the input CPAGs are transferred to the graph U. These two steps are the same as the first two steps of the ION algorithm. Figure 14 shows what the graph U looks like if we use the same input CPAGs as in the example from figure 9.

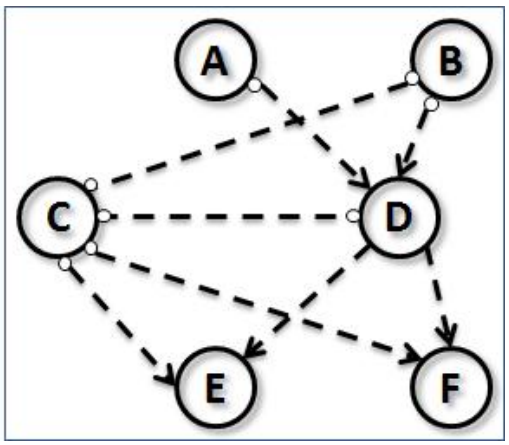

Figure 14 - Graph U (cSAT+ algorithm)

For every input CPAG a set of consistent MAGs is generated. Then, for every edge between two vari-

ables in graph U the cSAT+ algorithm checks if there is also an edge between those two variables in the corresponding set of MAGs. If this edge is present in all the corresponding MAGs, this edge is marked as a solid edge in graph U. If this edge is only present in one of the corresponding MAGs, this edge becomes a dashed edge in graph U. If this edge is not present in the corresponding set of MAGs, this edge is removed from the graph U. After this, every endpoint orientation is checked. If there is an edge between variable *X* and variable *Y* with an arrowhead on the side of variable *Y* and this is also the case in the set of corresponding MAGs, this arrowhead is added in graph U. If there is no arrowhead on the side of variable *Y* on the edge between variable *X* and variable *Y* in the corresponding set of MAGs, the arrowhead is placed on the side of variable *X* in graph U. If there are no more endpoint orientations to be checked, the graph U is returned as output, this is the PCG. Figure 15 shows what the output PCG looks like for this example.

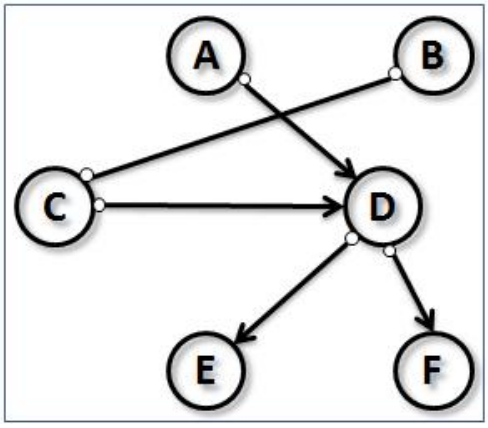

Figure 15 - Output PCG (cSAT+ algorithm)

9. CASE STUDY

As an example, I analyzed three different datasets containing data about ADHD (attention deficit hyperactivity disorder) patients. The first dataset contains ADHD data about children (Cao, 2009), the second dataset contains ADHD data about adolescents (von Rhein, 2014), and the third dataset contains ADHD data about adults (Hoogman, 2012). The data in the childhood and in the adulthood datasets is fMRI data and the data in the adolescence dataset is MRI data. These three datasets do not contain the exact same variables, but they do have an overlap in the variables they contain. I have selected seven variables from these three datasets to create an example with: gender, age, AD, HI, aggression, IQ, and DAT1. The adolescence dataset does not contain information about the variable age, so this variable is left out in this dataset. The adulthood dataset does not contain information about the variable aggression, so this variable is left out in this dataset. Also, the adulthood dataset is corrected for the variable gender, so unfortunately I cannot use gender as a variable for this dataset. Instead, information about the gene DAT1 is used in this dataset, this is a gene that has a proven effect on ADHD (Hoogman, 2012). Background information was also added to all of the three datasets: nothing can cause gender, age, or DAT1. Figures 16, 17, and 18 show the CPAGs that represent the causal relationships between the seven variables of interest for each of these three different datasets.

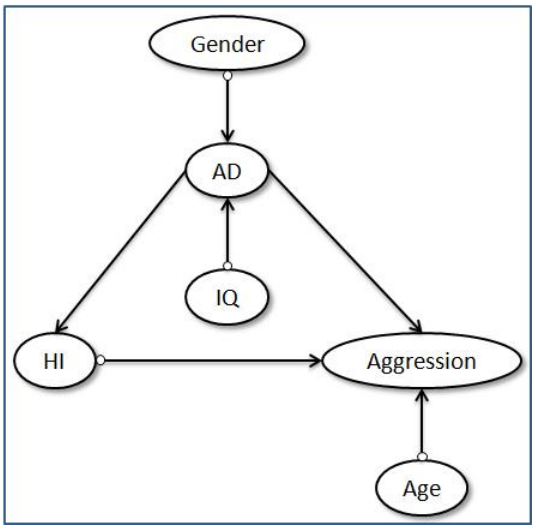

Figure 16 - CPAG from the ADHD childhood dataset

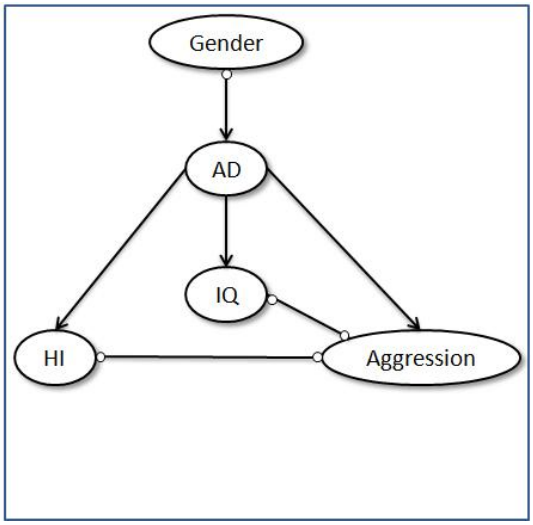

Figure 17 - CPAG from the ADHD adolescence dataset

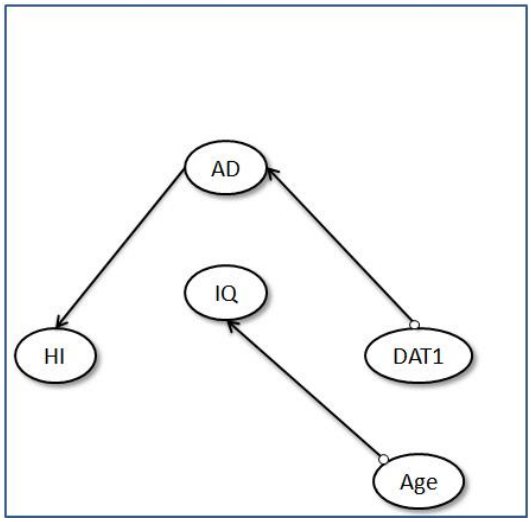

Figure 18 - CPAG from the ADHD adulthood dataset

In the first CPAG (from the childhood dataset) there is a causal link between the variable AD and the variable IQ, where IQ has a causal effect on AD. In the second CPAG (from the adolescence dataset) there is also a causal link between the variable AD and the variable IQ, but IQ does not have a causal effect on AD, instead, AD has a causal effect on IQ. This is conflicting information. The MCI-algorithm, the ION algorithm, and the cSAT+ algorithm cannot deal with conflicting information (R. Tillman, 2009), so these methods cannot be used to combine these three CPAGs. I proposed another method to try to deal with this situation. This method is an extension of the MCI-algorithm in combination with the BCCD algorithm. The MCI-algorithm is an algorithm that can only say if two variables have a causal relationship or not. The BCCD (Bayesian Constraint-based Causal Discovery) algorithm (Claassen T. &., 2012) calculates the probability of two variables having a causal relationship.

9.1. PROCEDURE

The three CPAGs are generated with the BCCD algorithm. For each of the three datasets the BCCD algorithm also calculates a table containing probability information about all the different causal relationships between the seven variables of interest. Those three different tables are put together in one table. If there is more than one probability for a particular causal relationship the relationship with the lowest probability is removed from the table. After that, I added the causal relationships one by one, from high probability to low probability, to the seven variables of interest. If there is a contradiction, I will stop adding the causal relationships and the CPAG that remains is the resulting CPAG. In this way, there will be no conflicting information in the resulting CPAG anymore.

9.2. RESULTS

Figure 19 shows the resulting CPAG from this method.

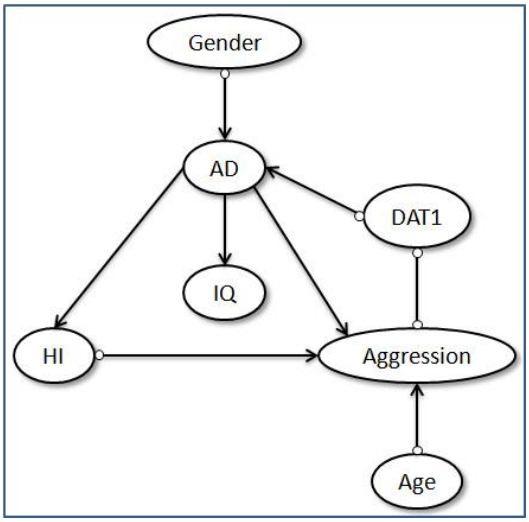

Figure 19 – First resulting CPAG

In this resulting causal model a few things stand out. The first thing that stands out is the relationship between DAT1 and aggression. This relationship should not be possible because these two variables do not occur together in the same dataset. Aggression is present in the childhood dataset and in the adolescence dataset and DAT1 is only present in the adulthood dataset. So, there is no data available about the possible relationship between aggression and DAT1. There is no statement in the probability table that says if there is an edge or if there is not an edge between aggression and DAT1 and that is why this edge is still present in the resulting causal model. But you cannot know for sure if this edge is really there or not, so to deal with this situation I represented this link as a dashed edge (figure 20). Also, I had to deal with the conflicting information in the childhood dataset and the adolescence dataset between variable AD and variable IQ. I looked at the probabilities of the contradicting causal relationships. The probability of AD influencing IQ in the adolescence dataset is 0,7789. The probability of AD not influencing IQ in the childhood dataset is 0,7127. This is not a big difference in probability, but I chose to pick the causal relationship with the highest probability (figure 20). So, the initial resulting causal model is not totally correct. I created a new resulting causal model with the

additions I made. Figure 20 shows this final resulting causal model.

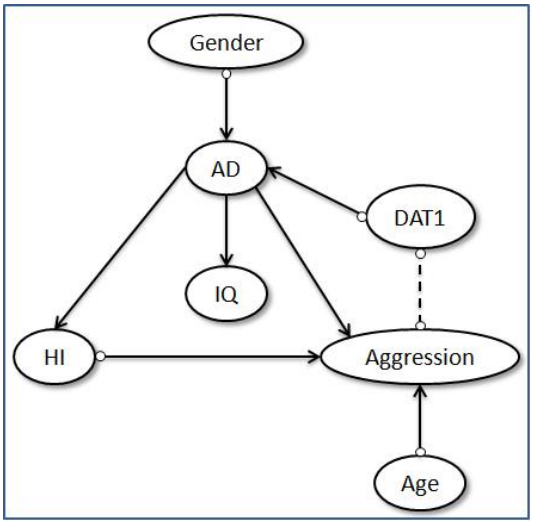

Figure 20 – Final resulting CPAG

9.3. PITFALLS

This does not mean that this final resulting CPAG is the only possible resulting causal model. There is no definitive final causal model. The childhood CPAG and the adolescence CPAG contain contradicting information, there could be three reasons for this: something went wrong in the algorithm that calculated the CPAG from the childhood dataset, the three different domains cannot be compared to each other (children, adolescents, and adults), and this method works with probabilities, so there is a chance that something can go wrong.

These three domains may not be suitable to be combined into one causal model. For example, boys have a higher chance on AD than girls, so gender has a high influence on AD in the childhood dataset and in the adolescence dataset. When these boys grow up, gender does not have a big effect on AD anymore. That is why it becomes hard to compare and combine these three different domains into one causal model. Also, the strength of, for example, the causal relationship between gender and AD can vary between the three datasets. So, the structure of the causal models stays the same, but the strength of the same causal relationship can vary. The method that I proposed, in combination with the BCCD algorithm, works, but only for datasets that have comparable domains and with some additions afterwards, like the dashed edge. This addition afterwards is the case if the variables of the different datasets are not exactly the same but have an overlap. In this case study, I made the assumption that the three domains of children, adolescents and adults were comparable, but it turned out that they might not to be comparable. I also chose to only add the causal relationship from variable age to variable aggression and to leave the causal relationship from variable age to variable IQ out because of the lower probability score. But I do not know exactly for sure that the causal relationship from variable age to variable IQ needed to be removed. So, this is another issue in working with this method. You cannot always be completely sure if a particular causal relationship should be included or not.

10. DIFFERENCES IN METHODS TO COMBINE CAUSAL MODELS

The MCI-algorithm, the ION algorithm, the IOD algorithm, and the cSAT+ algorithm all have the same goal: combining causal models from several different experiments. But these four methods all work in a different way and they thus have several differences when you compare them to each other. In this chapter these four different methods to combine causal models from different experiments will be compared to each other.

These are the differences that will be discussed here:
- Output
- Number of variables
- Conflicting information
- Speed

10.1. OUTPUT

The MCI-algorithm has a very clear and easily interpretable output in the form of one causal model (Claassen, 2010). This is also the case for the cSAT+ algorithm, the output is in the form of one pair-wise causal graph (Triantafilou, 2010). This is not the case for the ION algorithm and the IOD algorithm, the output of these two algorithms is in the form of a set of CPAGs (R. Tillman, 2009) and (Tillman R. E., 2011). So, the output of the MCI-algorithm and the cSAT+ algorithm is more clear and easier to deal with than the output of the ION algorithm and the IOD algorithm. If you compare the example output of the MCI-algorithm in figure 9 to the example output of the cSAT+ algorithm in figure 15, you can see that the edges of these two causal models are the same but there are three more edges in the causal model from the MCI-algorithm. This is because of the fact that the MCI-algorithm also takes the external environment into account as a set of unobserved, hidden variables. Because of this, there are more edges left in the output causal model of the MCI-algorithm than in the output model of the cSAT+ algorithm. The cSAT+ algorithm does not take the unobserved variables into account. For example, when there is a hidden common cause between variable *X* and variable *Y*, this will be recognized by the MCI-algorithm and there will be and edge between variable *X* and variable *Y* with arrowheads on both ends of this edge (a bi-directed edge). The cSAT+ algorithm will not recognize this hidden common cause and because of this there will be no edge between variable *X* and variable *Y*.

10.2. NUMBER OF VARIABLES

The MCI-algorithm can work with causal models that consist up to 20 nodes, this is not the case with the ION algorithm. The ION algorithm does not work well with many variables of interest (R. Tillman, 2009). The number of variables of interest also has an effect on the output of the ION algorithm. The more variables of interest there are, the more likely it is that the output CPAG set will be larger. This makes it harder to deal with the output. The number of variables of interest also has an effect on the

cSAT+ algorithm. The more variables of interest there are, the more time and memory the cSAT+ algorithm requires.

10.3. CONFLICTING INFORMATION

The MCI-algorithm, the ION algorithm, and the cSAT+ algorithm cannot deal with conflicting information (contradictions). The IOD algorithm can deal with conflicting information because of the testing for conditional independence and assigning p-values to pairs of variables in the first step of the IOD algorithm (Tillman R. E., 2011).

10.4. SPEED

Both the IOD algorithm and the cSAT+ algorithm are faster than the ION algorithm and require less memory (Triantafilou, 2010). This is because of the fact that the ION algorithm is a computationally hard algorithm and a brute force algorithm, which takes up a lot of time and a lot of memory.

11. CONCLUSION

At the moment, there are several different methods available to create causal models from the variables of interest in a single dataset (Tillman R. E., 2014). But in practice it often occurs that the different variables of two or more datasets have to be combined, including datasets that do not measure all the variables of interest, but have an overlap with the variables of the other datasets. Combining the variables of these different datasets is not possible with the methods that are used to work with a single dataset. It is very difficult to find or create a good working and reliable method to combine causal models from several different experiments to get a clear and more informative overview of the results. There are a few different methods available at the moment that can combine the causal models from different experiments. It is very useful to work with such a method because then there will be one clear causal model that contains all the causal information about all the variables of interest from all the different datasets, instead of multiple different causal models that have to be compared to each other.

I have discussed four of these methods that can combine causal models from several different experiments and compared them to each other: the MCI-algorithm, the ION algorithm, the IOD algorithm and the cSAT+ algorithm. They all have the same goal: combining the causal information from several different datasets. But they all work in a different way.

The MCI-algorithm and the cSAT+ algorithm produce the clearest output. Both methods give an output in the form of one causal model, which is easily interpretable. The ION algorithm and the IOD algorithm both give a larger output. Both methods give an output in the form of a set of CPAGs. These output sets can become very large and that makes it harder to deal with the output.

The MCI-algorithm can deal with a larger number of variables of interest than the ION algorithm and the cSAT+ algorithm. The more variables of interest there are, the more computation time and memory it costs with the ION algorithm and the cSAT+ algorithm.

The IOD algorithm can deal with conflicting information. This is not the case with the MCI-algorithm, the ION algorithm, and the cSAT+ algorithm.

The ION algorithm is not a very fast algorithm. The IOD algorithm and the cSAT+ algorithm are both faster algorithms than the ION algorithm.

On theoretical grounds, the MCI-algorithm is the best algorithm if you want to use an algorithm that produces a clear output and can handle a larger number of variables of interest. If you want to use an algorithm that can deal with conflicting information and is fast, you should use the IOD algorithm. This is all theoretically speaking, because in practice these methods all have their own pitfalls, so I could not use any of these methods in my case study. That is why I suggested another method.

The method that I used to combine the three CPAGs from the ADHD example datasets is a clear and simple method which can handle conflicting information, but this method only works for domains that can be compared to each other and with some additions afterwards.

REFERENCES

Borboudakis, G., & Tsamardinos, I. (2012). Incorporating causal prior knowledge as path-constraints in bayesian networks and maximal ancestral graphs. *ArXiv preprint arXiv: 1206.6390.*

Cao, X., Cao, Q., Long, X., Sun, L., Sui, M., Zhu, C., … & Wang, Y. (2009). Abnormal resting-state functional connectivity patterns of the putamen in medication-naïve children with attention deficit hyperactivity disorder. *Brain research*, 1303, 195-206.

Claassen, T., & Heskes, T. (2012). A Bayesian approach to constraint based causal inference. *arXiv preprint arXiv:1210.4866*.

Claassen, T., & Heskes, T. (2010). Causal discovery in multiple models from different experiments. In *Advances in Neural Information Processing Systems*, 415-423.

Claassen, T., & Heskes, T. (2010). Learning causal network structure from multiple (in) dependence models. In *Proc. Of the Fifth European Workshop on Probabilistic Graphical Models (PGM)*, 81-88.

Daly, R., Shen, Q., & Aitken, S. (2011). Learning Bayesian networks: approaches and issues. *The Knowledge Engineering Review*, 26(02), 99-157.

Danks, D., Glymour, C., & Tillman, R. E. (2009). Integrating locally learned causal structures with overlapping variables. In *Advances in Neural Information Processing Systems*, 1665-1672.

Geiger, D., Verma, T., & Pearl, J. (1990). Identifying independence in Bayesian networks. *Networks*, *20*(5), 507-534.

Grant, R. A. (1991). Building and testing a causal model of an information technology's impact. *Journal of Information Technology Management*, *2*(1), 11-23.

Hoogman, M., Onnink, M., Cools, R., Aarts, E., Kan, C., Arias Vasquez, A., … & Franke, B. (2013). The dopamine transporter haplotype and reward-related striatal responses in adult ADHD. *European Neuropsychopharmacology, 23(6),* 469-478.

Kleinberg, S., & Hripcsak, G. (2011). A review of causal inference for biomedical informatics. *Journal of biomedical informatics*, 44(6), 1102-1112.

Le, P. P., Bahl, A., & Ungar, L.H. (2004). Using prior knowledge to improve genetic network reconstruction from microarray data. *In Silico Biology 4*(3), 335-353.

Magidson, J. (1982). Some common pitfalls in causal analysis of categorical data. *Journal of Marketing Research* , 461-471.

Neapolitan, R. E. (2007, August 12). *Learning Bayesian Networks.* Retrieved January 21, 2015, from Videolectures.net: http://videolectures.net/kdd07_neapolitan_lbn/

Pearl, J. (1995). Causal diagrams for empirical research. *Biometrika*, *82*(4), 669-688.

Ramsey, J. D., Hanson, S. J., Hanson, C., Halchenko, Y. O., Poldrack, R. A., & Glymour, C. (2010). Six problems for causal inference from fMRI. *Neuroimage*, *49*(2), 1545-1558.

von Rhein, D., Mennes, M., van Ewijk, H., Groenman, A. P., Zwiers, M. P., Oosterlaan, J., … & Buitelaar, J. (2014). The NeuroIMAGE study: a prospective phenotypic, cognitive, genetic and MRI study in children with attention-deficit/hyperactivity disorder. Design and descriptives. *European child & adolescent psychiatry, 1-17*.

Richardson, T., & Spirtes, P. (2002). Ancestral graph Markov models. *Annals of Statistics*, 962-1030.

Russo, F. (2011). Explaining causal modelling. Or, what a causal model ought to explain. *New Essays in Logic and Philosophy of Science, College Publications, London.*

Sillignakis, K. E. (2001) The relative advantages and disadvantages of the causal and non-causal approaches to tourism demand forecasting.

Steel, D. (2011). Causality, causal models, and social mechanisms. *The SAGE Handbook of the Philosophy of Social Sciences*, 288-304.

Tillman, R. E., & Eberhardt, F. (2014). Learning causal structure from multiple datasets with similar variable sets. *Behaviormetrika, 41*(1), 41-64.

Tillman, R. E., & Spirtes, P. (2011). Learning equivalence classes of acyclic models with latent and selection variables from multiple datasets with overlapping variables. In *International Conference on Artificial Intelligence and Statistics*, 3-15.

Triantafilou, S., Tsamardinos, I., & Tollis, I. G. (2010). Learning causal structure from overlapping variable sets. In *International Conference on Artificial Intelligence and Statistics*, 860-867.

Tsamardinos, I., Triantafillou, S., & Lagani, V. (2012). Towards integrative causal analysis of heterogeneous data sets and studies. *The Journal of Machine Learning Research*, *13*(1), 1097-1157.

Zhang, J. (2008). On the completeness of orientation rules for causal discovery in the presence of latent confounders and selection bias. *Artificial Intelligence*, 172(16), 1873-1896.