New network models for the analysis of disease interaction

with applications in multimorbidity

Proefschrift

Ter verkrijging van de graad van doctor
aan de Radboud Universiteit Nijmegen
op gezag van de rector magnificus prof. mr. S.C.J.J. Kortmann,
volgens besluit van het college van decanen
in het openbaar te verdedigen op
maandag 2 juni 2014 om 16:30 uur precies
door

Martijn Lappenschaar

geboren 3 juni 1970 te Borculo.

Promotor: Prof. dr. P.J.F. Lucas

Copromotor: Dr. A.J. Hommersom

Manuscriptcommissie:

Prof. dr. T. Heskes – Radboud University
Prof. dr. S. Andreassen – Aalborg University
Prof. dr. M.G.M. Olde Rikkert – Radboud University Medical Center

Printed and cover layout by: Proefschriftmaken.nl, Uitgeverij BOXPress
Published by: Uitgeverij BOXPress, 's-Hertogenbosch

SIKS Dissertation Series No. XXX
The research reported in this thesis has been carried out under the auspices of SIKS, the Dutch Research School for Information and Knowledge Systems.

This research was funded by The Netherlands Organisation for Health Research and Development (ZonMw). Data acquisition took place at the Netherlands Institute for Health Services Research (NIVEL).

CONTENTS

1 INTRODUCTION 1
1.1 Interactions in nature 1
1.2 Correlation and causality 3
1.3 Clinical studies 4
1.4 Comorbidity and multimorbidity 5
1.5 Outline of this thesis 7
2 PROBABILISTIC GRAPHICAL MODELS 11
2.1 Probability theory 11
2.2 Bayesian networks 13
2.2.1 Definition 13
2.2.2 Parameter learning 14
2.2.3 Comparing models 15
2.2.4 Structure learning 16
2.3 Qualitative probabilistic networks 18
2.4 Chain graphs 19
2.5 Special cases of probabilistic graphical networks 20
2.5.1 Regression equations 20
2.5.2 Naive Bayes 21
2.5.3 Latent variable models 22
3 MULTIMORBIDITY IN GENERAL PRACTICE 25
3.1 Introduction 25
3.2 The definition of chronic diseases 26
3.3 Chronic diseases and clinical guidelines 26
3.4 Data collection in general practice 28
3.5 Disease counts of chronic diseases 30
3.6 Pairwise associations between chronic diseases 33
3.7 Clustering chronic diseases 36
3.8 Modelling effects of chronic diseases 39
3.9 Disease networks 41
3.10 Discussion 44
4 PROBABILISTIC CAUSAL MODELS OF MULTIMORBIDITY CONCEPTS 47
4.1 Introduction 47
4.2 Background 48
4.2.1 Comorbidity and multimorbidity 48
4.2.2 Causal relations within multimorbidity 49
4.3 Methods 50
4.4 Results 51
4.4.1 Existing comorbidity and multimorbidity concepts 51
4.4.2 Aetiological probabilistic models of multimorbidity 52

4.4.3 Probabilistic models for reasoning about clinical impact of multimorbidity 55
4.5 Discussion 57
5 FINDING CRITICAL FACTORS IN DISEASE CO-OCCURRENCE 61
5.1 Introduction 61
5.2 Structure learning of Bayesian networks and large datasets 62
5.2.1 Algorithms and data 62
5.2.2 Related work 63
5.2.3 Comparing Bayesian network structures 64
5.2.4 Limitations of structure learning 64
5.2.5 Structure learning from a large patient dataset 66
5.3 Models of co-occurrence and their characterisation 69
5.3.1 Statistical measures of association in multimorbidity 69
5.3.2 Structural measures of multimorbidity 71
5.3.3 Critical factors 71
5.3.4 Experiments 72
5.4 Conclusions 73
6 MULTILEVEL BAYESIAN NETWORKS FOR THE ANALYSIS OF HIERARCHICAL HEALTHCARE DATA 79
6.1 Introduction 79
6.2 Related research 80
6.3 Multilevel regression 82
6.4 Dealing with multilevel data by Bayesian networks 85
6.4.1 Basic ideas 85
6.4.2 Probability distributions for multilevel Bayesian networks 88
6.5 Experimental methodology 89
6.5.1 Parameter learning 89
6.5.2 Model validation 90
6.5.3 Structure learning 90
6.5.4 Artificial multimorbidity example with synthetic data 91
6.6 Modelling inter-practice variation in multimorbidity 95
6.6.1 Description of the models 95
6.6.2 Research problem and data 96
6.6.3 Unstructured MBNs compared to multilevel regression 96
6.6.4 Composition of the structured MBN 96
6.6.5 Comparison of the structured MBN with multilevel regression 98
6.7 Discussion 101
7 MULTILEVEL TEMPORAL BAYESIAN NETWORKS CAN MODEL LONGITUDINAL CHANGE IN MULTIMORBIDITY 105
7.1 Introduction 105
7.2 Methods 107
7.2.1 Data collection 107
7.2.2 Statistical analyses 107
7.3 Results 110

7.4 Discussion 114
7.4.1 Evaluation of the network structure 116
7.4.2 Quantitative analysis 116
7.4.3 Strength, limitations and implications 119
7.5 Conclusions 120
8 QUALITATIVE CHAIN GRAPHS AND THEIR APPLICATION 123
8.1 Introduction 123
8.2 Motivation from the medical field 125
8.3 Preliminaries 127
8.3.1 Notation 127
8.3.2 Chain graphs 128
8.3.3 QPNs 130
8.4 Qualitative chain graphs 132
8.4.1 Influences in chain graphs 132
8.4.2 Qualitative influences 134
8.4.3 Additive synergies 138
8.4.4 Intercausal reasoning and product synergies 139
8.5 Sign propagation 143
8.6 Experimental results 145
8.7 Conclusions 150
9 DISCUSSION 153
9.1 Main contributions 153
9.1.1 Bayesian networks as means to capture interactions 153
9.1.2 Bayesian networks for multilevel analysis 154
9.1.3 Chain graphs as means to model feedback systems 154
9.1.4 Multimorbidity and disease interactions 155
9.2 Future research 156
9.2.1 Probabilistic graphical models 156
9.2.2 Multimorbidity and mortality 156
9.3 Final note 157
A APPENDIX 161

Bibliography 170

Summary 189

Samenvatting 189

Dankwoord 191

Curriculum Vitae 193

1

INTRODUCTION

Many phenomena in the real world are a consequence of the interaction between entities of identical or different kind, whether at subatomic, atomic, molecular, cellular level, or at the level of people and populations of people. From a computing science point of view, quite naturally network models come to mind as a means to describe the interaction between such entities. Since the interaction between entities may have a stochastic nature, network models often have a probabilistic semantics.

The question that motivated the research described in this thesis was how we can make probabilistic networks practically useful for the detection of interactions between diseases – our entities here –, where some of these interactions might be novel, whereas others were already known. As people can have multiple diseases at the same time, network models can be useful to obtain insight into the interaction between these diseases, and the new probabilistic network methods in this thesis are therefore applied to the field of medicine, exploiting large datasets coming from primary care.

Discovering disease network models involves applying machine-learning methods to available data. The chosen setting of the research underlying this thesis, namely to exploit available large clinical datasets not especially gathered for research purposes, had major implications for the methods we had to develop. The idea of exploiting facts for understanding nature as it is, has been the basis of the empirical sciences at least since Galileo Galilei [59]. The importance of observations was also acknowledged by one of the most important scientists in the life sciences in the history of mankind:

> *My mind seems to have become a kind of machine for grinding general laws out of large collections of facts – Charles Darwin*

Nowadays many machine-learning techniques are available to distil models out of data. Many of these are also applicable to medicine, which led to the development of, e.g., probabilistic models that deal with the uncertainty related to diagnosis, prognosis, or the evaluation of therapy [116].

1.1 INTERACTIONS IN NATURE

Since ancient history, humans have tried to understand how one thing affects another. Sometimes the interaction between different entities are nowadays well understood and have been empirically verified. For example, the ideal gas law describes how the state of the amount of gas is determined by its pressure, volume, and temperature. Keeping

one of these variables constant, one obtains an isobaric, isochoric, or isothermal process, respectively. In these processes the interaction between the remaining variables is either directly or inversely proportional. For example, in an isobaric process, the interaction between volume and temperature is directly proportional, meaning that if the volume rises, the temperature also rises. Altogether, the interactions between pressure, volume, and temperature can be captured in one elegant mathematical model, which was first stated by Clapeyron in 1834. Recently, it was shown that the dynamics can also be modelled using a probabilistic network model [34].

As long as we can keep the amount of gas isolated from the outside world – a so-called *closed system* – the ideal gas law is a deterministic model. This would also imply that we cannot measure properties of the system, as then it would inevitably lose its closedness. Thus, open systems with unknown hidden factors are the norm, and we cannot predict exactly what will happen if we change one of the variables. At this point, probabilistic models come into sight. From domain knowledge we can build a probabilistic network model that can make predictions with a certain probability, and by using scientific evidence and observations we can make the model and its predictiveness more accurate [93].

Many interactions also have a *temporal dimension*, i.e., the effect can only be measured after a certain amount of time or it has a dynamic nature. It is in the people's nature trying to use the observations of today to predict what will happen in the future. Typical examples, and probably the most frequently used prediction models in the world, are those used for weather forecasting. Weather forecasts make use of current observations and experiences from the past, so the predictions are never entirely certain. Short term predictions are easier to make and more accurate, requiring fewer variables than predictions for the long term. The relevance of climate change for the latter type of predictions implies that many more variables with many more interactions between them need to be considered.

Probabilistic networks, e.g., Bayesian networks, have recently been used in this field of environmental modelling, providing a natural way to facilitate missing data, domain knowledge, and causal learning [202]. Their usefulness in making predictions about one of the major concerns in our world, the Arctic sea-ice loss and its interaction with greenhouse gas mitigation, has recently been demonstrated, including its influence on the polar bear population [4]. The loss of sea-ice extent in square miles is much more prominent in the summer and depends loosely on last year's extent and to a larger degree on the Arctic weather conditions, which in turn depends on the greenhouse effect. Figure 1.1 shows a very simplified 'plausible' model of the effect of human activities on the Arctic sea-ice extent and polar bear population.

Thus, there is growing evidence that probabilistic graphical models are practically useful in capturing interactions: they make it possible to connect all the domain variables involved in an intuitive network structure – with a temporal dimension if necessary – thereby allowing the researcher to investigate their interactions, qualitatively and quantitatively. Here in this thesis we apply them to the epidemiology of multiple chronic diseases that occur within one patient simultaneously, i.e., *multimorbidity*, to investigate the interactions between these diseases.

Probabilistic networks have been used in medicine before, but mainly from a single-disease perspective; networks containing multiple diseases were built for diagnostic

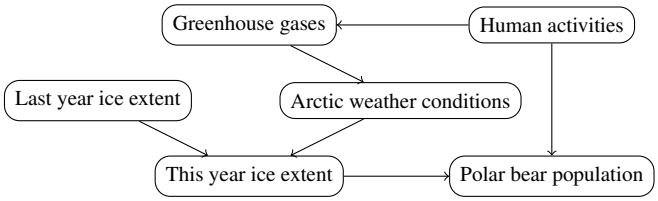


Figure 1.1: Simplified graphical model of the Arctic sea-ice extent and polar bear population. An edge between two entities means that the outcome of one influences the outcome of the other.

purposes [120]. So far, probabilistic models have not been used as a means to better understand the interaction among multiple chronic diseases. In particular, their usage with large medical *observational* datasets coming from multiple sources has not been explored.

1.2 CORRELATION AND CAUSALITY

"Correlation does not imply causation" is a phrase used in science and statistics to emphasise that a correlation between two variables does not necessarily imply that one causes the other. Many large datasets are observational and therefore not suitable for existing techniques that try to establish causality. Probabilistic graphical models, and in particular temporal models, can shed more light on cause and effect relationships in such datasets. Just recently, novel approaches using probabilistic graphical models were introduced to detect new causal relations – with a certain probability – from the combination of observations, even when coming from multiple different experiments [29].

However, to claim true causality remains a difficult task in medicine. For example, in the metabolic syndrome, lipid disorders, diabetes mellitus, and hypertension go side by side, and there are many hypotheses about the causal cascade of pathophysiologies. As another example, in case of a violent traumatic brain accident, one can still debate about the root cause. A psychiatrist might argue that we should not blame the perpetrator – who hit the patient's head – but that his or her behaviour is caused by a traumatic youth or a genetic predisposition. Darwin would argue that it is all part of the Evolution, and the true determinist originates everything back to the Big Bang.

David Hume – a Scottish philosopher in the 18^{th} century – took an even more sceptical position: he argued that knowledge about causality is based on experience, and experience is similarly based on the assumption that the future models the past, which in turn can only be based on experience, which leads to a circular logic. In conclusion, he asserted that causality is not based on actual reasoning, and that only correlation can actually be perceived. This emphasises that we should be cautious with conclusions drawn from clinical research that tries to address certain ground truth.

From a practical point of view it might sometimes be wise to accept a certain level of uncertainty about the true cause and just try to deal with the consequences. Both approaches are present in the history of medicine. For some conditions, there is a strong

focus on both determining the cause and finding ways to eliminate it; e.g., in cancer research both types of research are common. In other cases, the condition is accepted but its symptoms are treated using other physiological processes; e.g., use of diuretics for treating essential hypertension. The same distinction can be detected in psychiatry. Cognitive behavioural therapy tries to provide ways to cope with the psychiatric problems. On the other hand, there is a lot of research going on that tries to link psychiatric phenotypes, in particular developmental disorders, e.g., autism and attention deficit hyperactive disorders (ADHD), with biological substrates and genotypes.

If we can analyse how pathophysiological processes can be influenced – where elimination is (yet) not possible – this can help us to find ways for treating multiple diseases. If a condition A is (patho)physiologically linked to condition B, treating condition A also *affects* condition B. Secondly, if we can analyse the statistical influences over time between multiple diseases, this can help us to shed more light on causality. If a condition A is associated with a condition B later in time – under certain conditions known as Granger causality – this might imply direct causality. If indeed, we find such patterns in temporal observational data, the justification of a clinical study would gain more support.

1.3 CLINICAL STUDIES

Nowadays, many patient data is present in so-called electronic health records, which exist in information systems of hospitals, general practices, insurance companies, and so on. From a scientific point of view the nature of patient data present in electronic health records resembles to the one of observational studies. Because of the size of such datasets, its analysis may be taken as belonging to the 'Big Data' problem. Although their usage for treatment comparisons is not recommended due to treatment selection bias [23], they may be very useful in other research studies, e.g., in the areas of epidemiology, prognosis, diagnosis, quality improvement, and healthcare planning. But even for these purposes, it is often not clear how inference should be performed and in which way the right conclusions should be drawn [37].

The patient's health status with respect to multiple chronic diseases is a complex process. Both diagnosis and therapy become more biased each time a new disease emerges within the patient. The identification of all the interactions that play a role in this process can only be accomplished by using large sets of patient data. However, most of the clinical research is typically carried out by using randomised controlled trials (RCTs). Randomised experiments first appeared in psychology, where they were introduced by Charles Sanders Peirce in 1885 [152], and the first published RCT appeared in 1948 [130]. A brief history of the RCTs can be found in [131].

RCTs are designed to focus on one or two outcome variables using a relatively small patient dataset. They often exclude patients with comorbidity and polypharmacy, which makes them more or less unsuitable to investigate multiple diseases at the same time. On the other hand, when using observational studies, one should take notice of the bias introduced by possible non-random selection criteria. For example, several prominent medical researchers issued the Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) statement, in which they called for observational studies to

conform to twenty-two criteria that would make their conclusions easier to understand and generalise [49].

Since there are hardly any randomised clinical trials that cover a large set of chronic diseases, the focus of this thesis is on patient data retrieved from large observational studies, e.g., from general practices. Recent explorations of patient data from primary care registries to quantify associations between chronic disorders, have shown these to be valuable for obtaining a broad picture of chronic diseases [11, 77, 225].

1.4 COMORBIDITY AND MULTIMORBIDITY

The term *comorbidity* was introduced in 1970 by Feinstein [50]. He defined comorbidity as the occurrence of other medical conditions additional to an index disease. Before the term *multimorbidity* was coined, there were already several indices that measured the impact of co-existing diseases on the outcome of a disease variable related to the primary disease of interest. A user's guide to select such *comorbidity indices* for clinical research was presented by Hall in 2006 [71]. He discussed the four most commonly used general comorbidity indices: the cumulative illness rating scale (CIRS) [113], the Kaplan-Feinstein classification (KFC) [91], the Charlson comorbidity index (CCI) [27], and the index of co-existent disease (ICED) [30]. Other popular measures are the chronic disease score (CDS), the adjusted clinical groups (ACG) system, and the Duke severity illness checklist (DUSOI); all three are discussed in [85].

More recently, the term multimorbidity has been introduced in chronic disease epidemiology to refer to any co-occurrence of two but often more than two medical conditions within a person [206]. The introduction of this term indicates a shift of interest from a given index condition to the individuals who have multiple disorders. Figure 1.2 shows that in the last five years the usage of the term *multimorbidity* in scientific research has grown exponentionally. However, in a substantial part of this research, multimorbidity is used as an explaining variable in a regression model of a primary disease of interest. In these cases the term *comorbidity* would be a more appropriate one to use.

The same holds for the identified *reviews* in PubMed with the term multimorbidity in the Title or Abstract. Only a limited amount of them have a focus on multimorbidity without a perspective from a single disease or intervention; they are listed in Table 1.1. In particular, in the last three years there have been a tremendous effort to systematically review the multimorbidity related research papers. The European General Practice Research Network reports eleven main themes in multimorbidity research: chronic disease, acute disease, biopsychosocial factors and somatic risk factors, coping strategies of the patient, burden of the disease, healthcare consumption, disability, quality of life, frailty, social network, and health outcome.

A systematic review on *multimorbidity indices*, i.e., indices that address the severity of the patient's disease status with respect to the presence of multiple diseases, can be found in the work of Diederichs et al. [39]. They concluded that the literature further emphasises the heterogeneity of existing multimorbidity indices. However, one important similarity is that the focus is on diseases with a high prevalence and a severe impact on affected individuals.

Currently, multimorbidity research still has a strong focus on disease counts, pairwise associations and clustering methods. Recently, *disease networks* were proposed

Title of article	Year	Journal
Comorbidity or multimorbidity; what's in a name? A review of the literature [206]	1996	Eur J General Practice
Problems in determining occurrence rates of multimorbidity [208]	2001	J Clin Epidemiol
Multimorbidity is common to family practice: is it commonly researched? [52]	2005	Can Fam Physician
Defining comorbidity: implications for understanding health and health services [204]	2009	Ann Fam Med
The measurement of multiple chronic diseases – a systematic review on existing multimorbidity indices [39]	2011	J Gerontol A Biol Sic Med Sci
Aging with multimorbidity: a systematic review of the literature [125]	2011	Ageing Res Rev
Measures of multimorbidity and morbidity burden for use in primary care and community settings: a systematic review and guide [85]	2012	Ann Fam Med
A systematic review of prevalence studies on multimorbidity: toward a more uniform methodology [54]	2012	Ann Fam Med
Multimorbidity in primary care: a systematic review of prospective cohort studies [55]	2012	Br J Gen Pract
Managing patients with multimorbidity: systematic review of interventions in primary care and community settings [192]	2012	BMJ
Multimorbidity in Older Adults [178]	2013	Epid Rev
The European General Practice Research Network Presents a Comprehensive Definition of Multimorbidity in Family Medicine and Long Term Care, Following a Systematic Review of Relevant Literature [166]	2013	JAMDA

Table 1.1: Systematic reviews of multimorbidity.

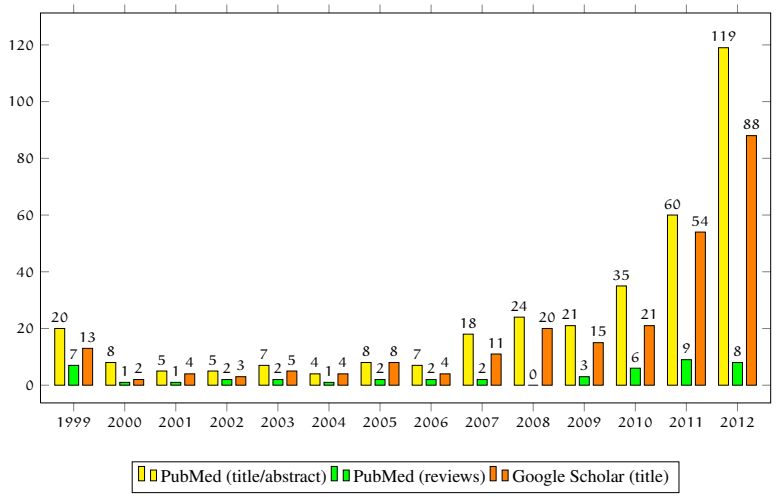


Figure 1.2: Multimorbidity as search term in PubMed and Google Scholar.

as a method for modelling disease progression of multiple diseases and have found their way in both somatic diseases [10] and psychiatric diseases [33]. However, the interactions in such disease networks are still defined based on pairwise associations rather than on conditionally independence, something which is provided in probabilistic graphical models. A key advantage of using a probabilistic network for the representation of uncertain knowledge is that they offer an easily understandable graphical statistical model of how disease variables interact with each other. In our research we adopted them as the main technique for modelling multimorbidity.

1.5 OUTLINE OF THIS THESIS

With the ageing of the population in western countries comes the fact that many of the elderly are nowadays faced with the presence of multiple chronic diseases at the same time. Traditionally, diagnostic processes often try to find one disease that explains all the symptoms presented by the patient. Applying therapy to multiple diseases often involves mostly adding up what is recommended by the separate guidelines on individual diseases. In the future, an integrated approach would be more desirable to meet the individual needs of a patient that is faced with multiple diseases [69].

This goal has still a long way to go, and here we take a step forwards by determining how interactions between multiple chronic diseases should preferably be analysed from a epidemiological point of view. Starting from the statistics that currently exist in multimorbidity research, for example, total disease counts or pairwise associations, we move to graphical models in which diseases and their related measurements are put in a *disease network* where the connections (edges) represent their statistical dependences. By doing this, we ensure that every patient variable – and thus each disease already present – still matters, when zooming into the model for a particular disease. In short,

we try to bridge the gap that exists between current statistical methods and the demands resulting from an integral approach when facing a patient with multimorbidity.

In Chapter 2 we review the basic concepts of probabilistic graphical models (PGMs) as used in this thesis, i.e., Bayesian networks, qualitative probabilistic networks, and chain graphs. Besides that, we also summarise the basic properties of the more traditional techniques used, i.e., regression analysis and principal component analysis, and show that they can be seen as special cases of PGMs.

In Chapter 3 we first provide background information about clinical guidelines and the patient data used in this thesis, followed by a short review of the current state of the art of methodologies used in current multimorbidity research. These methodologies can be divided into five main themes:

1. Total disease counts of chronic diseases.
2. Pairwise associations between chronic diseases.
3. Clustering of chronic diseases.
4. Effects of chronic diseases.
5. Disease networks.

We will evaluate and illustrate some of the techniques used in other research by applying them to a large patient dataset extracted from multiple general practices in the Netherlands. In this way, we avoid comparing apples and oranges when using all the multimorbidity research results obtained from the literature. Moreover, it can be shown that a certain caution should be taken when applying these techniques to a large observational dataset that includes multiple diagnosis.

In Chapter 4, we go back to the formal definitions of multimorbidity and show how probabilistic networks can be used to model most of the concepts used in multimorbidity. We summarise existing classifications and terminologies used in definitions related to multimorbidity and point out their similarities and differences. It turns out that many of the used terminologies are similar from a probabilistic point of view. We show that a limited number of probabilistic network templates can be used to formalise most of the existing terminologies used in multimorbidity research. This work was also presented in [102].

in Chapter 5 we introduce a statistical measure, called *critical factors*, that best explains the co-occurrence of two diseases when the complete set of disease variables is arranged in a Bayesian network. Learning the structure of such a network is a complex task, and we show that the outcome is fairly sensitive to the sample size and extreme prevalences. Having determined which structure learning technique is empirically the most useful in large sets of patient data, we learned the *critical factors* between pairwise combinations of malignant tumours. These critical factors can roughly be divided into patient characteristics, lifestyle related conditions, and pre-malignant pathophysiology. Furthermore, the resulting networks also reveal the (known) pathways of metastasis. This work was partially presented in [103].

In Chapter 6, we explore the fact that certain biases can be introduced when data is extracted from multiple patient datasets. The data we used was extracted from multiple general practices. It is known that some of these practices have a higher (or lower)

prevalence of certain chronic diseases because of population differences induced by the practice related variables, e.g., urbanity, region, and average age. This requires a multilevel analysis, e.g., multilevel regression, that takes into account this bias. Given the perceived higher usefulness of Bayesian networks for the analysis of multiple diseases, we introduced a new concept, called multilevel Bayesian networks (MBNs), that incorporates multilevel analysis into Bayesian network. The concept is illustrated with an generated dataset and with real-life patient data from over a hundred general practices in the Netherlands. This work appeared in [105].

In Chapter 7 we extend the MBN with a temporal dimension. The resulting framework is used to evaluate the associations between cardiovascular diseases and their progression over a period of five years. We used three major chronic health conditions: obesity, hypertension, and lipid disorders, and we analysed their effect on six (groups of) cardiovascular (related) diseases, i.e., diabetes mellitus, ischaemic heart disease, heart failure, stroke, retinopathy, and nephropathy. The same model is also used to evaluate the progression of total disease count over time. This work appeared in [106].

The idea of *qualitative influences*, as introduced in Chapter 2, is further studied in Chapter 8, where we explore the consequences of the fact that a one-way causal direction between two associated diseases cannot always be determined. Many physiological processes are maintained in a kind of equilibrium state, and pathophysiology can be seen as a disturbance of such equilibria. Previously, it was shown that chain graphs – which is a hybrid graph containing both directed and undirected edges – can model such equilibria [107, 34]. Using the existing properties of chain graphs and qualitative reasoning we introduced qualitative chain graphs (QCG) and explored its usefulness in the medical domain. This work is under revision in [104].

Finally, in Chapter 9, our results are discussed and put in a general context. We also provide recommendations for future research in the area of multimorbidity. The human body is a very complex system; it constantly tries to maintain homoeostasis and it has many physiological and behavioural adaptations to cope with chronic disturbances. Integrating more and more of these concepts into one model – the road to a patient oriented clinical guideline – makes inference of such a model a hard task. Moreover, numbers that explain some of the more rare conditions might be overwhelmed by those concerning more common chronic conditions. The hybrid solution of Chapter 8, which incorporates both expert knowledge and empirical knowledge into one model in a qualitative manner, provides a possible way to integrate knowledge from different sources.

> *Complex, statistically improbable things are by their nature more difficult to explain than simple, statistically probable things – Richard Dawkins*

2

PROBABILISTIC GRAPHICAL MODELS

ABSTRACT

In this chapter, we summarise the basic statistical concepts that are required to understand the remaining chapters. Particular details of a topic that are only dealt with in one specific chapter are discussed in that chapter itself. Probabilistic graphical models, e.g., Bayesian networks and chain graphs, are used to model the statistical dependences and independences that exist between observed variables. There are several techniques to learn both the structure and the parameters of such models, which will be described here. The *qualitative* abstraction of a Bayesian network is often referred to as a qualitative probabilistic network. In such a network we speak of influences and synergies rather than the *quantitative* differences in probabilities. Finally, commonly used statistical techniques, for example, regression and latent class analysis, can be viewed upon as special cases of probabilistic graphical models, and will also be discussed briefly.

2.1 PROBABILITY THEORY

We start with a brief summary of notation and some basic concepts. Random variables are denoted by upper case characters, e.g., X; a value is indicated by lower case characters or numbers, e.g., x. The expression $X = x$, with the equality predicate '$=$', is a logical expression that is either true or false. Other logical operators and predicates can also be used within expressions, depending on the nature of the random variable of concern. Often we will write x rather than $X = x$ to indicate that a variable X has value x. A random variable X can have one of the values from its domain, i.e., its values are mutually exclusive. In contract to variables, for random variables values are real numbers with associated properties, such as that they are totally ordered. Often, however, we will not make a distinction between variables and random variables.

In case variables are binary with values *true* and *false*, its values are denoted by x and $\overline{x}$ respectively. If the value of a variable is known, this is referred to as an *observation*, an *instantiation*, or as *evidence*.

A random variable X is always assumed to be a set of variables, being a singleton set when we are dealing with a single variable. We assume there is a multivariate, or joint, probability distribution over the set of random variables X, denoted by $P(X)$. The joint probability distribution of two sets X and Y is denoted by $P(X, Y)$. When the actual value of a random variable does not matter, we will often also write $P(X)$ rather

than $P(X = x)$ for a probability. We assume the reader is familiar with Kolmogorov's axioms of probability theory.

We make a distinction between discrete and continuous random variables. For the discrete case, the probability distribution can be defined by a probability *mass* function $f_X : \mathbb{R} \to [0, 1]$ and it holds that $P(X = x) = f_X(x)$. An associated distribution function F_X is defined in terms of the probability mass function f_X as follows: $F_X(x) = \sum_{-\infty}^{x} f_X(x)$. For the continuous case, a probability distribution is defined indirectly by a probability *density* function $f_X : \mathbb{R} \to [0, 1]$, such that the associated distribution function is defined as $F_X(x) = \int_{-\infty}^{x} f_X(u)du$. From the basic axioms of probability theory, it follows then for both discrete and continuous random variables that: $F_X(x) = P(X \leqslant x)$.

We now turn to some useful properties of joint probability distributions.

Definition 1 (conditioning)**.** *Let* P *be a joint probability distribution of a set of variables* X*. A conditional probability distribution* $P(X \mid Y)$ *is defined as:*

$$P(X, Y)/P(Y)$$

for positive $P(Y)$*. The corresponding conditional density or mass functions are denoted by* $f_{X|Y}$*.*

Proposition 1 (chain rule)**.** *Let* P *be a joint probability distribution of a set of variables* $X = \{X_1, \ldots, X_i, \ldots, X_n\}$*. It holds that:*

$$P(X_1, \ldots, X_n) = P(X_n \mid X_1, \ldots, X_{n-1}) \cdots P(X_2 \mid X_1)P(X_1)$$

Definition 2 (marginalisation)**.** *Let* P *be a joint probability distribution of a set of variables* X*. The marginal distribution of* $Y \subseteq X$ *for discrete variables is defined as:*

$$P(Y) = \sum_{Z=X\setminus Y} P(Y, Z)$$

Similarly, for continuous variables the marginal density function of Y *is defined as:*

$$f_Y(y) = \int_{z:Z=X\setminus Y} f_{Y,Z}(y, z)dz$$

with $f_{Y,Z}$ *the joint probability density function.*

Definition 3 (independence)**.** *Two sets of variables* X *and* Y *are said to be conditionally independent given a third set of variables* Z*, denoted as* $X \perp\!\!\!\perp_P Y \mid Z$*, if*

$$P(X \mid Y, Z) = P(X \mid Z)$$

for any value of Y*. If, in contrast, these variables are conditionally dependent, this is denoted by* $X \not\perp\!\!\!\perp_P Y \mid Z$*.*

Theorem 1 (Bayes' rule)**.** *Let* P *be a joint probability distribution of a set of variables* X*. Let* $Y, Z \in X$ *then:*

$$P(Y \mid Z) = \frac{P(Z \mid Y)P(Y)}{P(Z)}$$

where $P(Z) > 0$*.*

The *odds* is defined as the ratio of the probability that a particular event will happen to the probability that the event will not happen:

$$\mathrm{odds}(X = 1) = \frac{P(X = 1)}{P(X = 0)} = p/(1-p)$$

for a binary variable X with $P(X = 1) = p$.

In case the sequence of random variables $X = \{X_1, \ldots, X_n\}$ is a time series and they adhere to the Markov property, i.e., given the present state X_i, the future state X_{i+1}, and the past states $X_1, \ldots, X_{i-1}$ are independent, we speak of a Markov chain. More formally,

$$P(X_{i+1} \mid X_1, \ldots, X_i) = P(X_{i+1} \mid X_i)$$

Note that the possible values of X_i should form a countable set S, also called the state space of the Markov chain. Very often, Markov chains are described using a directed graph, where the edges represent the transitions between states. Such transitions occur with a certain probability, and the parameters of the probability distributions are often attached as labels to the edges.

2.2 BAYESIAN NETWORKS

2.2.1 *Definition*

Formally, a *Bayesian network*, BN for short, is a tuple $\mathcal{B} = (G, X, P)$, with $G = (V, E)$ a directed acyclic graph (DAG), $X = \{X_v \mid v \in V\}$ a set of random variables indexed by the set of vertices V, E a set of directed edges (also called arcs) between vertices in V, and P a joint probability distribution of the random variables in X. P is a Bayesian network with respect to the graph G if P can be written as a product of the probability of each random variable, conditional on their parent variables:

$$P(x_1, \ldots, x_n) = \prod_{v \in V} P(x_v \mid x_{pa(v)}) \tag{1}$$

where $pa(v)$ is the set of parents of v (i.e. those vertices pointing directly to v via a single arc). Likewise we have children of v, and together with the parents they are called the *neighbours* of a vertex, denoted by $ne(v)$. As a convenience, we will often write v if we mean the random variable X_v that is associated to v. In Bayesian networks, the arcs between variables model dependences between variables which give rise to probabilistic conditional independence relationships. The graph G is an *independency map* (I-map), which means that the independences implied by G also hold in P, i.e.:

$$X \perp\!\!\!\perp_G Y \mid Z \implies X \perp\!\!\!\perp_P Y \mid Z \tag{2}$$

where $X \perp\!\!\!\perp_G Y \mid Z$ can be read off the graph G using the well-known criterion of d-separation [149].

The graph where all arcs are replaced by undirected edges is called the *skeleton* of a Bayesian network. The *Markov blanket* (MB) of a vertex v is the set of vertices such that v is d-separated of all other vertices given the set of vertices in the Markov blanket.

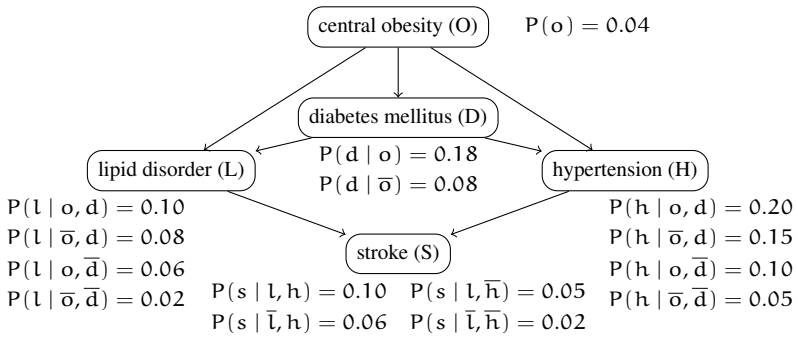


Figure 2.1: Example of a Bayesian network.

In a BN, the Markov blanket of a vertex is the set of parents, children, and parents of children.

Figure 2.1 shows an example of a Bayesian network. In this network, for example, it holds that $S \perp\!\!\!\perp O \mid L, H$. Furthermore, the Markov blanket of H, for example, is $\{O, D, L, S\}$, where the variables O, D are a parent, S is a child, and L is a parent of a child.

2.2.2 *Parameter learning*

Suppose we have a set of independent and identically distributed observations $x = (x_1, \ldots, x_n)$ and a statistical model $f(x; \theta)$ with parameters $\theta = (\theta_1, \ldots, \theta_k)$, that explains the observations. For an independent and identically distributed sample, the joint probability density function is:

$$f(x; \theta) = f(x_1; \theta) f(x_2; \theta) \ldots f(x_n; \theta) = \prod_{i=1}^{n} f(x_i; \theta) \tag{3}$$

with x varying freely and θ a fixed parameter. Now we look at this function from a different perspective by considering the observed values x as fixed parameters and θ as a variable. This results into the likelihood:

$$L(\theta) = \prod_{i=1}^{n} f(x_i; \theta) \tag{4}$$

Very often the log-likelihood is used, which turns the product sign into a summation:

$$\log L(\theta) = \sum_{i=1}^{n} \log f(x_i; \theta) \tag{5}$$

Let $L(\widehat{\theta})$ be the maximised value of the likelihood function of the model, i.e., $L(\widehat{\theta}) = \arg\max_{\theta} L(\theta)$. This defines a maximum likelihood estimator (MLE) of θ. Since the function log is monotonically increasing, the MLE estimate is the same regardless of whether we maximise Equation 4 or Equation 5.

For example, suppose X is a single binary variable with $P(X = 1) = \theta$ and $P(X = 0) = 1 - \theta$. Then $f(x;\theta) = \theta^x(1-\theta)^{1-x}$, and the log-likelihood is defined as:

$$\log L(\theta) = n_1 \log\theta + n_0 \log(1-\theta)$$

with n_1 the number of observations where $X = 1$, n_0 the number of observations where $X = 0$, and $n_1 + n_0 = n$. The MLE then becomes: $\widehat{\theta} = n_1/(n_1+n_0) = n_1/n$. In case X is a multi-valued discrete random variable with $P(X = i) = \theta_i$ and $i \in \{i_1, \ldots, i_k\}$, it is easy to prove that $\widehat{\theta}_i = n_i/n$, with n_i the number of observations where $X = i$.

In case of a Bayesian network the likelihood can be decomposed according to the decomposition in Equation 1:

$$L(\theta) = \prod_{v \in V} \prod_{i=1}^{n} f_{v|pa(v)}(x_{v,i};\theta) = \prod_{v \in V} L_v(\theta_v) \tag{6}$$

with $pa(v)$ the parents of X_v. For discrete variables the likelihood can be further decomposed resulting in $\widehat{\theta}_{v|pa(v)} = n_{v|pa(v)}/n_{pa(v)}$.

2.2.3 *Comparing models*

In case of competing models, there are several criteria to select the most desirable model. The most well-known approaches are Akaike's information criterion (AIC) [2], which is based on the Kullback-Leibler (K-L) information loss [99], and the Bayesian information criterion (BIC) [183], which is based on Bayesian factors.

Definition 4 (K-L information)**.** *Let* g *denotes the full reality or truth (*g *has no parameters), and let* h *be an approximating model with a probability distribution. The K-L information, denoted as* $I(g,h)$*, is the information lost when a model* h *is used to approximate* g*:*

$$I(g,h) = \int g(x) \log\left(\frac{g(x)}{h(x \mid \theta)}\right) dx \tag{7}$$

Since g does not depend on the data nor on the model, this can be rewritten as:

$$I(g,h) = C - E_g\left[\log(h(x \mid \theta))\right] \tag{8}$$

Replacing θ with the MLE $\widehat{\theta}$, this becomes:

$$I(g,h) = C - E_O E_g\left[\log(h(x \mid \widehat{\theta}))\right] \tag{9}$$

with O the set of observations. Akaike showed that the latter part can be estimated by the maximised log-likelihood value with a bias equal to the number of estimated parameters k in the model, i.e.,:

$$C - E_O E_g\left[\log(h(x \mid \widehat{\theta}))\right] = \log(L(\widehat{\theta})) - k \tag{10}$$

For historical reasons the right part is multiplied by -2, and this became Akaike's information criterion for model selection.

Definition 5 (AIC)**.**

$$\mathrm{AIC} = -2\log(L(\widehat{\theta})) + 2k \tag{11}$$

The model with the lowest AIC value is seen as the best model that approximates the truth. A few years after Akaike published his results, Schwarz proposed an alternative metric that puts a larger penalty on the number of parameters in the model:

Definition 6 (BIC)**.**

$$\mathrm{BIC} = -2\log(L(\widehat{\theta})) + \log(n)k \tag{12}$$

Nowadays, there are still debates on which one to choose, and of both metrics there are a number of variants [22].

2.2.4 *Structure learning*

In case one is not sure about the dependences and independences within the network structure G of a BN we can apply structure learning. Given a dataset D with N observations, i.e., $D = (D_1, \ldots, D_N)$ with D_i an instantiation of all the variables in V, Bayesian network structure learning is the problem of learning a network structure G from D. This can be done using various methods, which are basically divided into constraint-based methods, score-based (or search-and-score-based) methods, and hybrid methods.

The constraint-based methods make use of statistical independence tests to directly test whether variables are independent from each other. The score-based methods employ a measure that scores a network structure G given the data D, and use this measure to carry out a a heuristic search through the space of DAGs, and sometimes the space of equivalence classes of DAGs. As the number of possible graphs grows more than exponentionally with the number of variables, one uses mostly a greedy search algorithm to obtain the graph that minimises the score. The hybrid methods combine ideas from constraint-based and score-based methods.

In this thesis, we applied the following algorithms that were implemented in the *bnlearn R package* [185]: (*i*) the grow-shrink (GS) algorithm [126], (*ii*) the tabu search (TABU) algorithm [126], and (*iii*) the max-min hill-climbing (MMHC) algorithm [200]. We will describe them briefly here for the discrete case, following the definitions in [200, 115, 185, 126].

The GS algorithm is a constraint-based structure learning algorithm. It uses the Markov blanket information of vertices to determine the structure. The Markov blankets are determined by a grow and a shrink phase as follows:

1. $\mathrm{MB}(v) \leftarrow \emptyset$
2. Growing phase: while $\exists w \in V - \{v\}$ such that $w \not\perp\!\!\!\perp v \mid \mathrm{MB}(v)$, do: $\mathrm{MB}(v) \leftarrow \mathrm{MB}(v) \cup \{w\}$.
3. Shrinking phase: while $\exists w \in \mathrm{MB}(v)$ such that $w \perp\!\!\!\perp v \mid \mathrm{MB}(v) - \{w\}$, do: $\mathrm{MB}(v) \leftarrow \mathrm{MB}(v) - \{w\}$.

The network structure is then determined as follows:

1. Determine the direct neighbours of all the vertices: $\forall v \in V, w \in \mathrm{MB}(v) : \forall S \subseteq T : v \not\perp\!\!\!\perp w \mid S \rightarrow w \in \mathrm{ne}(v)$, where T is the smallest of the sets $\mathrm{MB}(v) - \{w\}$ and $\mathrm{MB}(w) - \{v\}$.

2. Determine which direct neighbours are a parent: $\forall v \in V, w \in \mathrm{ne}(v) : \exists u \in \mathrm{ne}(v) - \mathrm{ne}(w) - v$, such that $\forall S \subseteq T : w \not\perp\!\!\!\perp u \mid S \cup \{v\} \rightarrow w \in \mathrm{pa}(v)$, where T is the smallest of the sets $\mathrm{MB}(v) - \{u, w\}$ and $\mathrm{MB}(u) - \{v, w\}$.

3. Remove those cycles by identifying the minimal set of edges that need to be reversed for all cycles to disappear.

4. Propagate remaining directions: $\forall v \in V, w \in \mathrm{ne}(v)$ such that $w \rightarrow v \notin G$ and $v \rightarrow w \notin G$, do: if there exists a directed path from v to w, orient $v \rightarrow w$.

Tabu search is a score-based structure learning algorithm. In these algorithms, a scoring function is used to measure the goodness of fit of a learned structure. The score approximates the probability of the structure given the data and represents a trade-off between how well the network fits the data and how complex the network is. Assuming that the scoring function is *decomposable*, the score for a Bayesian network structure G can be calculated as the sum of scores for individual variables:

$$\mathrm{Score}(G \mid D) = \sum_{v \in V} \mathrm{Score}(v \mid \mathrm{pa}(v), D) \tag{13}$$

The learning problem is then to find G^* such that $G^* = \arg\max_G \mathrm{Score}(G \mid D)$.

Usually, the scoring functions are in the form of a penalised log-likelihood function. If we take the MLE $\widehat{\theta}$ for the parameters θ given the graph G, it holds that:

$$P(D \mid G) = P(D \mid G, \widehat{\theta}) \tag{14}$$

where:

$$\log P(D \mid G, \widehat{\theta}) = \sum_{v \in V} \sum_{i=1}^{N} \log P(v_i \mid \mathrm{pa}(v_i)) \tag{15}$$

with v_i and $\mathrm{pa}(v_i)$ instantiations in data point D_i. Note that Equation 15 is similar to the log of Equation 6, except that here $P(D \mid G, \widehat{\theta})$ is a function of G, rather than θ. Then, the penalised LL is defined as:

$$\mathrm{LL}_{\mathrm{pen}}(G, D) = \log P(D \mid G) - \sum_{v \in V} \mathrm{Penalty}(v, G, D) \tag{16}$$

It is easy to see that this penalised LL is decomposable as in Equation 13. In this thesis, three different penalties are used:

1. $\mathrm{Penalty}_{\mathrm{AIC}}(v, G, D) = p_v$,

2. $\mathrm{Penalty}_{\mathrm{BIC}}(v, G, D) = p_v \cdot \frac{1}{2} \log N$

3. $\mathrm{Penalty}_{\mathrm{BDE}}(v, G, D) = \sum_{j \in \mathrm{pa}(v)} \sum_{k=1}^{K_v} \log \frac{P(D_{ijk} \mid D_{ij})}{P(D_{ijk} \mid D_{ij}, \alpha)}$

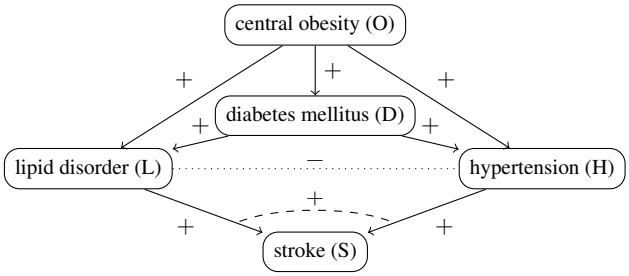


Figure 2.2: Example of a qualitative probabilistic network.

with p_ν the number of parameters for ν, K_ν the number of possible values for ν, D_{ijk} the number of times $\nu_i = k$ and $pa(\nu_i) = j$ in D. The parameter α is a hyperparameter, called the *equivalent sample size*, to make scores equal for networks within the same equivalence class [74]. The penalties 1 and 2 are equivalent to those used in Definition 5 (AIC) and 6 (BIC).

Because the number of possible structures is more than exponential in the number of variables [171], TABU employs a local search method to limit the search. The MMHC Bayesian network learning algorithm is a hybrid algorithm. The algorithm first identifies the parents and children set of each variable, then performs a greedy hill-climbing search in the space of Bayesian networks. The search begins with an empty graph. The edge addition, deletion, or direction reversal that leads to the largest increase in score is taken and the search continues in a similar fashion recursively. The same scores as used in TABU can be used for this step in the MMHC algorithm.

2.3 QUALITATIVE PROBABILISTIC NETWORKS

Qualitative probabilistic networks (QPNs) were introduced by Wellman [221], as a qualitative abstraction of Bayesian networks. Conditional probability distributions are replaced by qualitative knowledge in the form of signs, which describe the relationships among variables by the concepts of probabilistic influences and synergies.

A *qualitative influence* expresses how the value of one variable influences the probability of observing values of another variable. In addition to influences, a qualitative probabilistic network includes synergies modelling interactions between influences. An *additive synergy* expresses how the interaction between two variables influences the probability of observing the values of a third variable. *Product synergies* are used to provide *intercausal reasoning*, i.e., they express how upon observation of a common child of two vertices, observing the value of one parent vertex influences the probability of observing a value of the other parent.

Consider the qualitative network in Figure 2.2, which is the same network as in Figure 2.1, but now with qualitative signs instead of probability distributions. The plus signs along the edges represent positive influences. For example, there is a positive influence of hypertension on stroke, meaning that if hypertension is present there is a higher probability of stroke, no matter the value of lipid disorder. The plus sign above the dashed line represents a positive synergy between hypertension and lipid disorder

on stroke, meaning that there is an interaction between hypertension and lipid disorder, making stroke more likely. On the other hand, once it is *known* that stroke is present, the probability of one condition (either hypertension or lipid disorder) decreases in the presence of the other – it is so-called 'explained away' – which is represented by the minus sign above the dotted line.

More formally, we say that A has a *positive qualitative influence on* B if

$$P(b \mid a, Z) \geqslant P(b \mid \overline{a}, Z) \tag{17}$$

with Z the set of variables $\mathrm{pa}(B) \setminus \{A\}$. A *negative* influence, and a *zero* influence, are defined analogously, by replacing $\geqslant$ with $\leqslant$ and $=$ respectively. If none of this holds, the influence is called *ambiguous*.

Secondly, we say there is a *positive additive synergy* of A_1 and A_2 on B if

$$P(b \mid a_1, a_2, Z) + P(b \mid \overline{a}_1, \overline{a}_2, Z) \geqslant P(b \mid \overline{a}_1, a_2, Z) + P(b \mid a_1, \overline{a}_2, Z) \tag{18}$$

with Z the set consisting of the variables $\mathrm{pa}(B) \setminus \{A_1, A_2\}$.

Finally, we say there is a *positive product synergy* of A_1 and A_2 with regard to the value b of variable B if

$$P(b \mid a_1, a_2, Z) \cdot P(b \mid \overline{a}_1, \overline{a}_2, Z) \geqslant P(b \mid \overline{a}_1, a_2, Z) \cdot P(b \mid a_1, \overline{a}_2, Z) \tag{19}$$

Negative, *zero*, and *ambiguous* additive and product synergies are defined analogously.

Both influences and synergies adhere to a set of convenient properties, such as symmetry [221], which will be further discussed in Chapter 8.

2.4 CHAIN GRAPHS

In a Bayesian network each edge is directed, which implies that for each individual vertex a probability distribution can be defined that depends only on the vertex's parents. In chain graphs, however, some of the edges are undirected, which implies that probability distributions are defined over cliques of vertices rather than for the individual vertices. In the abstraction where the cliques of vertices are replaced by single, composite vertices, the resulting network resembles again a Bayesian network, in fact a polytree-shaped BN. This property is captured by the outer factorisation of the definition of the joint probability distribution, which follows later.

To illustrate the concept of a chain graph, Figure 2.3 shows a possible chain graph over the variables used in Figure 2.1, along with its Bayesian network abstraction. In the chain graph the edges between diabetes mellitus, lipid disorder and hypertension are undirected. The probability distribution of these variables together with obesity is defined by the potentials φ_{ODHL}, which puts ratios on the probabilities within this clique of vertices. The probability of stroke is conditioned by diabetes mellitus, lipid disorder, and hypertension only. In the Bayesian network abstraction, the latter diseases are represented by a single vertex.

More formally, associated to a chain graph $G = (V, E)$ is a joint probability distribution P over the set of vertices V that is faithful to the chain graph G, i.e., it contains

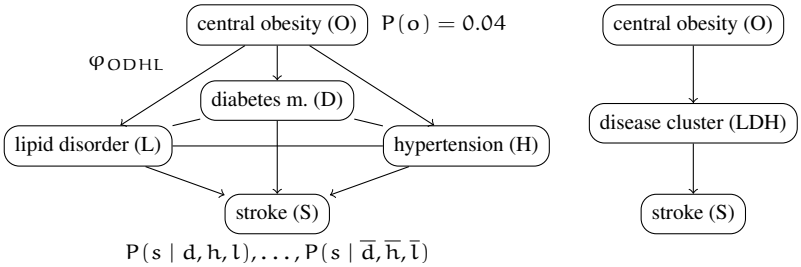


Figure 2.3: Example of a chain graph (on the left) and its Bayesian abstraction – representing the outer factorisation of the joint probability distribution – on the right.

all the independences implied by the graph. Such distributions can be factorised by an *outer factorisation* of the cliques in $\mathcal{C}$:

$$P(V) = \prod_{C \in \mathcal{C}} P(C \mid pa(C)) \tag{20}$$

with the set of vertices $V = \bigcup_{C \in \mathcal{C}} C$, and where each $P(C \mid pa(C))$ is defined by a clique-wise factorisation:

$$P(C \mid pa(C)) = Z^{-1}(pa(C)) \prod_{M \in M_C} \varphi_M(M) \tag{21}$$

given that M_C are the complete (fully connected) subsets in the *closure* graph of C, i.e., the subgraph $G_{C \cup pa(C)}$ where each arc is replaced by a line and each pair of vertices of $pa(C)$ is also connected by a line, also referred to as moralization. The functions φ are non-negative real functions, called *potentials*; they generalise joint probability distributions in the sense that they do not need to be normalised. Finally, the normalising factor Z is defined as:

$$Z(pa(C)) = \sum_{C} \prod_{M \in M_C} \varphi_M(M) \tag{22}$$

If a chain graph only contains directed edges, then it is a Bayesian networks, as then each chain component consists of a single vertex, so only the outer factorisation applies. Chain graphs that only contain undirected edges are called *Markov networks*. More details on chain graphs are provided in Chapter 8.

2.5 SPECIAL CASES OF PROBABILISTIC GRAPHICAL NETWORKS

2.5.1 *Regression equations*

In a Bayesian network in which there are solely edges from the explanatory variables to the outcome variable, i.e., $P(O, E_1, \ldots, E_n) = P(O \mid E_1, \ldots, E_n) \prod_{i=1}^{n} P(E_i)$, with O the outcome variable and E_i the explanatory variables, the conditional probability distribution $P(O \mid E_1, \ldots, E_n)$ can be estimated using a regression model. For

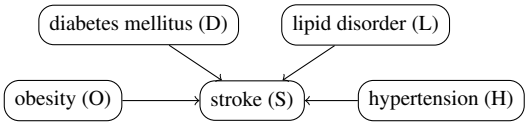


Figure 2.4: Example of regression modelled as a Bayesian network .

example, in Figure 2.4 the outcome variable stroke can be explained by regression of the variables obesity, diabetes mellitus, lipid disorder, and hypertension.

In general it is assumed that the outcome variable O is linear dependent with weights β on the set of explanations e through a link function g, i.e., $g(E[O|e]) = \beta^T e$. This is the so-called generalised linear model (GLM). For example, in the standard *linear* regression model, the link function is defined as $g(x) = x$, and the regression model becomes:

$$P(O \mid e) \sim \text{Normal}(\mu, \sigma) \tag{23a}$$
$$\mu = \beta^T e \tag{23b}$$

With $e = (1, e_1, \ldots, e_i, \ldots, e_n)^T$ and $\beta = (\beta_0, \beta_1, \ldots, \beta_n)$. In case of binary outcome variables, the *logistic* regression model is often used. In this case, the link function g is given by the logit function, i.e., $g(x) = \text{logit}\, x = \frac{x}{1-x}$, and the model becomes:

$$P(O \mid e) \sim \text{Bernoulli}(p) \tag{24a}$$
$$\text{logit}\, p = \beta^T e \tag{24b}$$

In case the outcome variable is a count variable, e.g., the total disease count, the *Poisson* or *negative binomial* regression models are often used. Their link function is given by the log function, i.e., $g(x) = \log x$, and the model becomes:

$$P(O \mid e) \sim \text{Poisson}(\lambda) \text{ or } \text{Negbin}(\lambda, \tau) \tag{25a}$$
$$\log \lambda = \beta^T e \tag{25b}$$

Regression models can be extended to deal with data that is hierarchically organised. This will be further discussed in Chapter 6.

2.5.2 *Naive Bayes*

Graphically, the naive Bayes model is the reverse model compared to the regression model. The edges now point to the explanations, see e.g., Figure 2.5, and the Bayesian network decomposition becomes:

$$P(O, E_1, \ldots, E_n) = P(O) \prod_{i=1}^{n} P(E_i \mid O) \tag{26a}$$

Just like in regression analysis, the computation of the parameters in the naive Bayes model is tractable and its predictive power is comparable to logistic regression. It was

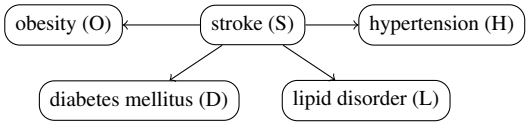


Figure 2.5: Example of a naive Bayes model.

shown that naive Bayes reaches its asymptotic error very quickly with regards to the number of training examples. Thus, if training data is scarce, one can expect naive Bayes to outperform logistic regression, but as the number of training examples grows, logistic regression will outperform naive Bayes and achieve a lower asymptotic error rate [137].

2.5.3 *Latent variable models*

In certain modelling techniques it is assumed that the observed association between observed variables is caused by a set of latent (hidden) variables [12]. Given a set of specific values for such variables, the observed variables are then independent, e.g., see Figure 2.6. The joint probability distribution of the set of observed variables $O = \{O_1, \ldots, O_n\}$ and the set of latent variables L can be factorised as follows:

$$P(O, L) = P(L) \prod_{i=1}^{n} P(O_i \mid L) \tag{27}$$

In case of factor analysis the (conditional) distribution of both the observed and latent variables are considered to be Gaussian:

$$P(O_i \mid L) \sim \text{Normal}(\alpha_i + \beta_i L, \sigma) \tag{28a}$$

$$P(L) \sim \text{Normal}(0, 1) \tag{28b}$$

In a latent class analysis (LCA) the associations between observations are also clarified by means of hidden (latent) variables. But here the latent variables are considered to be discrete.

$$P(O_i \mid L) \sim F_i(L) \tag{29a}$$

$$P(L) \sim \text{Categorical}(\phi) \tag{29b}$$

Theoretically, the conditional distribution F_i can follow any distribution, but in most parameter estimators it is considered to be either Gaussian, Poisson, or categorical.

In factor analysis, one can distinguish between an explanatory factor analysis (EFA) and a confirmatory factor analysis (CFA). In an EFA, the researcher's a priori assumption is that any observed variable may be associated with any latent variable. There is no prior theory and one uses factor loadings to intuit the structure of the data. A CFA tries to determine if the number of latent variables and the loadings of observed variables on them, is conform to what is expected on the basis of a pre-established theory. The researcher's a priori assumption is that each latent variable is associated with a specified subset of observed variables. In a CFA, a number of analyses can be used to

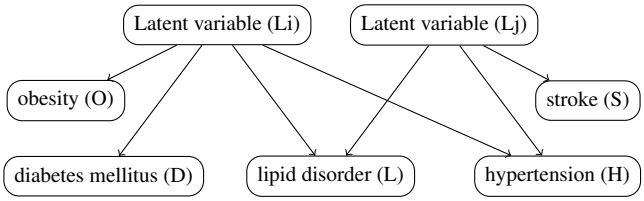


Figure 2.6: Example of a model with latent variables.

determine if the model has a good fit, whereas in an EFA one tries to find the optimal number of latent variables that best explains the set of observations [12].

The number of latent classes in LCA can be obtained by comparing fit indices of solutions with varying numbers of classes. Typical fit indices for LCA are the likelihood ratio χ^2 statistic, the Akaike Information Criterion (AIC) and the Bayesian information criterion (BIC). Additionally, the Lo-Mendell-Rubin adjusted likelihood ratio test can be used to determine the number of latent classes in the model [143]. It allows one to compare the model with k classes to the model with $k-1$ classes. Based on the p-value of the significance test one can choose to reject the model with k classes and turn back to the model with $k-1$ classes.

The intuition behind the difference in determining the parameters of an LCA model is that factor analysis is concerned with clustering variables, whereas LCA is more concerned with clustering cases, i.e., the latent taxonomic structure. Therefore, LCA is more analogously closer to cluster analysis, e.g., item response theory (IRT) and grade-of-membership (GOM) analysis [123].

3

MULTIMORBIDITY IN GENERAL PRACTICE

ABSTRACT

There are many diseases that put a chronic burden on the patient's quality of life, especially when they co-occur in one patient. Collections of large observational patient datasets are frequently used to examine the epidemiology of multiple chronic diseases occurring at the same time. The resulting multimorbidity figures can be divided into five main themes: total disease counts, pairwise associations, clusters of diseases, temporal effects of multimorbidity, and disease networks. Here we discuss their properties and illustrate them using a large patient dataset, which was obtained from multiple general practices in the Netherlands. Except for childhood and ages above eighty, total disease counts were proportional with age on a logarithmic scale. Pairwise combinations of diseases often show a higher prevalence than expected, especially at higher ages. The more complex figures obtained with cluster analysis and disease networks show that diabetes mellitus, lipid disorders, hypertension, and musculoskeletal disorders, are responsible for many of the observed associations between other chronic diseases. Using a Bayesian network, which takes into account these confounding effects, the resulting network of the most commonly observed chronic diseases is less dense and more comprehensive than the one obtained by taking all the significant pairwise associations.

3.1 INTRODUCTION

Epidemiological research indicates that more than two third of the elderly have two or more chronic diseases at the same time; this problem, one of the most challenging of modern medicine, is referred to as the problem of comorbidity or multimorbidity. Its focus has been increasing lately, and, as pointed out in Chapter 1, a large number of multimorbidity indices are available these days. In this chapter, we provide more background information about several aspects of multimorbidity in primary care. First, we describe how and when diseases are considered to be a chronic disease. Next, we provide the necessary information about the patient data that was used for analysis throughout this thesis. This patient data was collected from multiple general practices in the Netherlands.

In the sections thereafter, we review the most popular techniques that are currently used to analyse multimorbidity in general practice. These techniques are illustrated by

the same patient data used further in this thesis. Finally, we briefly discuss the pros and cons of these techniques and demonstrate a first preliminary use of a probabilistic graphical network concerning multimorbidity in general practice.

3.2 THE DEFINITION OF CHRONIC DISEASES

In the definition of multimorbidity, a disease is seen as a 'chronic disease' if it is irreversible without any expectation of complete recovery and with a relatively long period of illness or recurrence of illness. Patients with chronic diseases distinguish themselves by a prolonged need of healthcare. This definition is further detailed by O'Halloran et al. [144]. The complete list (last accessed on October 26, 2012) of chronic diseases made by O'Halloran, based on the international classification of primary care (ICPC) codes, can be found at:

- http://sydney.edu.au/medicine/fmrc/classifications/DefiningChronicConditions.pdf

Based on these criteria, a definition based on the ICPC codes was provided in 2008 for the Netherlands [79], which divided chronic diseases into 29 disease groups. This list (last accessed on October 26, 2012) can be found at:

- http://www.nationaalkompas.nl/gezondheid-en-ziekte/ziekten-en-aandoeningen/chronische-ziekten-en-multimorbiditeit/selectie-van-chronische-ziekten/

This led to a final set of chronic diseases used for analysis in this thesis, which is listed in Table 3.1. The chronic diseases are organised in thirteen groups, following the main morphology of the human body. For the detailed specification of the subgroups, e.g., 'not otherwise specified' (NOS) or 'other', we refer to [144].

3.3 CHRONIC DISEASES AND CLINICAL GUIDELINES

Clinical guidelines provide recommendations to physicians and patients, mostly about the management of one disease; the recommendations are based as much as possible on scientific evidence. The diagnostic process of a disease is supported by indicating relevant symptoms, risk factors, signs obtained by physical examination and laboratory investigations to decrease the uncertainty in the diagnosis. Treatments are supported by providing advice through a stepwise approach with regular control of the patient's symptoms and signs, until a proper balance in the patient's condition is achieved.

In case of the elderly, the physician often deals with multimorbidity, and then several guidelines need to be consulted. As medical knowledge in clinical guidelines is organised around single disorders, this knowledge may not be fully applicable to patients with multiple disorders [17], offering no guarantees that elderly patients receive appropriate treatment. Eventually this can lead to conflicts, e.g., treatment of one disease can reduce the efficacy of a treatment of another disease or worsen its outcome. It has been noted that guidelines should be adapted to account for multimorbidity [69], and multimorbidity has not only implications for healthcare, but also for research and

ICPC chapter	Chronic disease
A – General/Infections	tuberculosis, malignant neoplasm, congenital anomaly NOS, weakness or tiredness syndromes: chronic fatigue syndrome, myalgic encephalomyelitis, post viral fatigue syndrome;
B – Haematology	Hodgkin and non-Hodgkin lymphoma, leukaemia, malignant neoplasm NOS, benign neoplasm, hereditary haemolytic anaemia, vitamin B12 and folate deficiency anaemia, anaemia NOS, purpura and coagulation defects, HIV, AIDS;
D – Gastroenterology	viral hepatitis, malignant neoplasm: stomach, colon, rectum, pancreas, other locations; congenital anomaly, duodenal ulcer, peptic ulcer, diverticular disease, irritable bowel syndrome, chronic enteritis, ulcerative colitis, liver disease NOS, cholecystitis, cholelithiasis, other: e.g., pancreatitis, gluten and lactose intolerance, stenosis;
F - Ophthalmology	retinopathy, macular degeneration, cataract, glaucoma, blindness, neoplasm;
H - Otolaryngology	vertiginous syndrome, presbyacusis, deafness, neoplasm;
K - Cardiovascular	rheumatic fever, neoplasm, congenital anomaly, ischaemic heart disease with and without angina pectoris, acute myocardial infarction, heart failure, atrial fibrillation/flutter, paroxysmal tachycardia, cardiac arrhythmia NOS , heart murmur NOS, pulmonary heart disease, heart valve disease NOS, heart disease NOS, hypertension uncomplicated/complicate, postural hypotension, transient cerebral ischaemia, stroke, cerebrovascular accident, cerebrovascular disease, atherosclerosis, peripheral vascular disease, pulmonary embolism, phlebitis, thrombophlebitis, varicose veins of leg;
L - Musculoskeletal	malignant neoplasm, congenital anomaly, neck syndrome, back syndrome with and without radiating pain, acquired spine deformity, rheumatoid and seropositive arthritis, osteoarthritis: hip, knee, other locations; shoulder syndrome, tennis elbow, osteoporosis, other: e.g., osteitis, polymyositis, dystrophy;
N - Neurology	neurologic infection, malignant neoplasm, benign neoplasm, congenital anomaly, multiple sclerosis, parkinsonism, epilepsy, migraine, cluster headache, trigeminal neuralgia, carpal tunnel syndrome, peripheral neuritis and neuropathy, other: e.g., encephalopathy, palsy, paraplegia;
P - Psychiatry	chronic alcohol abuse, dementia, schizophrenia, affective psychosis, organic psychosis, anxiety disorder, somatisation disorder, depressive disorder, neurasthenia, compulsive disorder, personality disorder, hyperkinetic disorder, post-traumatic stress disorder, mental retardation, anorexia nervosa, anorexia bulimia, psychosis NOS;
R - Pulmonology	malignant neoplasm: bronchus, lung, other locations; hypertrophy tonsils, adenoids, chronic obstructive pulmonary disease, asthma, other: e.g., asbestosis, fibrosis;
S - Dermatology	malignant neoplasm, seborrhoeic dermatitis, dermatitis and atopic eczema, psoriasis, acne, other: e.g., rosacea, lupus, lichen sclerosis;
T - Endocrinology	malignant neoplasm thyroid, neoplasm other, congenital anomaly, goitre, obesity, overweight, hyperthyroidism, hypothyroidism, diabetes: insulin and non-insulin dependent; gout, lipid disorder, other: e.g., Cushing, Addison, cystic fibrosis, haemochromatosis;
U - Urology	malignant neoplasm: kidney, bladder, other; glomerulonephritis, nephrosis, other: chronic renal failure or insufficiency, urethral stenosis;

Table 3.1: Chronic diseases (NOS=not otherwise specified).

medical education [11]. Eventually, there is a need for an electronic guideline that can be easily adapted to each patient [16, 84].

With the continuous advancement in healthcare due to scientific research, clinical guidelines are constantly adjusted by incorporating the newest scientific evidence. For example, for a long while, hypertension guidelines focused on blood pressure as the only or main parameter determining the need and the type of treatment. Modern guidelines on hypertension now emphasise that the diagnosis and management of hypertension should be based on the quantification of total cardiovascular risk. Therefore, it also establishes connections with other chronic diseases. For example, the presence of obesity, diabetes, or lipid disorders, worsen the prognosis of hypertension and common consequences of hypertension are cerebral vascular disease, heart disease, renal disease and retinopathy.

There are many ways in which computing science can play a role in improving clinical guidelines. In related research, clinical guidelines, as computer-program-like textual documents, were formalised and verified mathematically using temporal logics [82]. In this thesis, the emphasis is on using probabilistic methods to capture the uncertain relationships between diseases, based on data, and to reasoning about the presence and evolution of diseases by means of probabilistic reasoning methods.

3.4 DATA COLLECTION IN GENERAL PRACTICE

The patient data used for analysis were obtained from the register of the Netherlands Information Network of General Practice (LINH). It started in 1996 as a register of referrals of general practitioners to medical specialists. Nowadays, twice a year, information about contacts and diagnoses, prescriptions, referrals and – since 2007 – lab and physiological measurements are extracted from the practice information systems. All Dutch inhabitants are obligatory registered with a general practice, and the LINH database contains information of routinely recorded data from about all patients of approximately 90 general practices. Only practices with a proper registration were used in the analysis.

The first Dutch General Practitioners Information Systems, GPIS for short, date from the 1980's. These systems allow for making and recording appointments with patients, on the one hand, and recording all other patient-related and care-related information a doctor is supposed to collect, on the other hand. There are around eight different vendors of GPIS in the Netherlands, some of whom have their origin in the pharmacy IT. The names of these different GPIS are: Medicom, Promedico, MicroHIS, OmniHis, Mira, Webhis Zorgdossier, HetHis and TetraHis, ordered from Medicom, which has the largest market share, to TetraHis having the smallest. LINH is a representative network with respect to the population, type of practice and type of GPIS. The first five of the above mentioned GPIS are supported by the LINH database. For each GPIS, software was developed to extract coded data from the systems.

The Dutch college of General Practitioners (NHG) develops the so-called GPIS referential model. Every GPIS should support all features and entities described in this model. For this purpose, the ICT companies developing the GPIS are subscribed to receive annual updates of this model. Despite the fact that every type of GPIS has its

own design and, therefore, database structure, LINH pre-processes the extracted data in such a way that it fits the uniform LINH data model.

Before data are stored in the LINH database, several validity checks are performed. First, it is checked whether a received file contains data and to which practice the data belongs. In addition, the number of registered weeks is checked to see whether data is missing, and the order of variables listed is checked, as variables could be swapped, for example because of an update of the system to a new release at the general practice. Also, the average number of records per patient are calculated and compared to developed standards.

In the LINH relational database model, data are patient-centred and, thus, a record can always be related to a patient, and to a specific date of the event that took place in a specific GP practice. For example, laboratory measurements or prescriptions for a patient have an attached date. Every module (prescription, referral, consult, etcetera) is represented in a table and has a unique identifier. Every patient has a unique identifier, the client id. Every time new data is extracted, the data are linked to the same patients who are already known to the database before adding all these data incrementally to the database. The client table stores all patient information, such as to which practice the patient belongs. The practice table has been excluded from the model; it contains information on the location, and thus the degree of urbanisation, the number of GPs, and the GPIS the practice uses. When data are stored in the MS SQL database, type-checking is performed on the data.

As a last step, quality checks on the data registered by the GP are performed to, for example, calculate the number of ICPC codes recorded, ATC coded prescriptions, etcetera. These checks are done on the basis of criteria developed by the LINH research team. Only good quality data are used for LINH research. Every GP practice receives feedback on the delivered dataset.

The dataset used in this thesis is a subset from the LINH dataset, consisting of consultations, measurements, referrals, and prescriptions until December 2011. Figure 3.1 shows the relational data-model of this particular dataset, showing that the data is hierarchically organised by one-to-many relationships.

Having obtained the data, we noticed that laboratory results and medication were not always consistent with the diagnoses present in the LINH database. For example, insulin was sometimes prescribed to patient who were not diagnosed with diabetes mellitus according to the data. To compensate for such missing or incorrect information, we used lab results and medication to infer the diagnosis of obesity, hypertension, lipid disorders, and diabetes mellitus, in conjunction to the ICPC codes.

Corrections were made by using the following rules adopted from the Dutch guideline on cardiovascular risk management [223], which are in line with the European guidelines on cardiovascular risk management in clinical practice [156]. For obesity: a body mass index over 30 kg/m^2; for hypertension: a high blood pressure (systolic $>$ 140 mm Hg or diastolic $>$ 90 mm Hg) within at least two recurring measurements; for lipid disorders: an abnormal blood lipid profile (low density lipoprotein $>$ 3 mmol/l, high density lipoprotein $<$ 1 mmol/l, or triglycerides $>$ 2 mmol/l); and for diabetes mellitus: a fasting glucose $\geqslant$ 6 mmol/l, or a prescription of either insulin or oral blood glucose lowering medication.

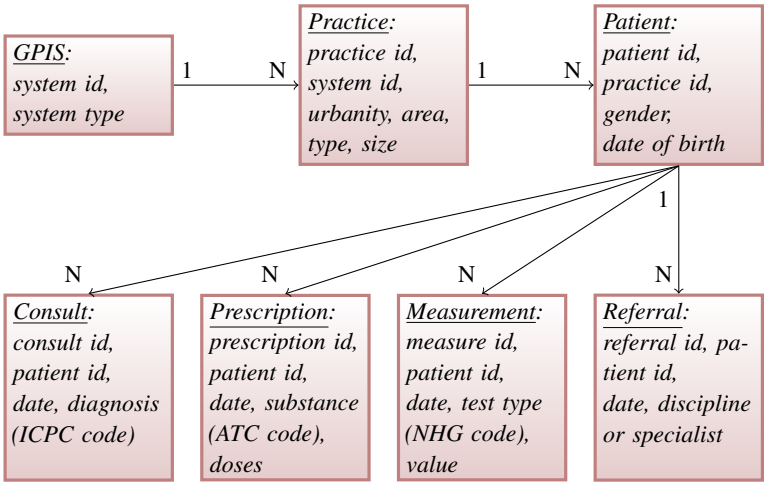


Figure 3.1: High level model of the patient data retrieved from general practices in the Netherlands.

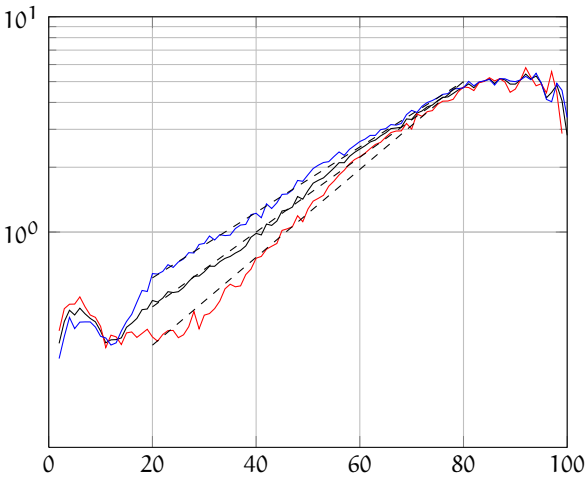
3.5 DISEASE COUNTS OF CHRONIC DISEASES

The easiest way to measure multimorbidity is to count the number of diseases per patient, and determine the average number of diseases by age, see e.g., [224, 53, 201]. In this way, Van den Akker et al. [207] already determined prevalences and indices of multimorbidity (defined as two or more co-occurring diseases) for the Dutch population in the nineties of the last century. Overall, the ratio of the observed and the expected number of diseases had a U-shaped distribution when plotted for age. The differences between the observed and expected number of diseases were strongly statistically significant ($p < 0.0001$) for both sexes and all ages.

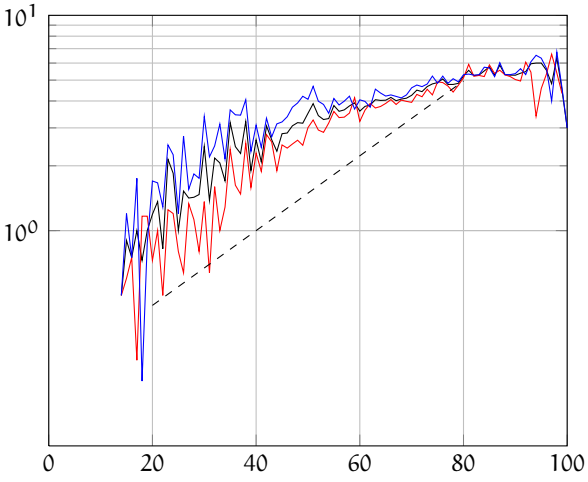
When modelling count data, i.e., a discrete variable taking values $0, 1, 2, ..., K$, with lower values more likely, *Poisson* regression is often the first choice. Alternatively, one can use the *geometric* or *negative binomial* distribution. Real-life data frequently display overdispersion and excess zeros. The former regression methods are then extended with an additional parameter leading to *zero-inflated* Poisson, geometric, or negative binomial regression. Most of the research on multimorbidity starts with modelling disease counts using one of these techniques, see e.g., [157, 6, 193, 48].

Using the *lme4* package in the statistical software package *R* [1], we applied the count models, as described above, to the LINH patient dataset. Table 3.2 shows the resulting Akaike information criterion (AIC) and Bayesian information criterion (BIC), for all patients and a specific patient group, i.e., diabetics. The negative binomial zero-inflated models show the best fit, and Figure 3.2 shows the observed disease counts by age and gender on a logarithmic scale, along with the estimations derived by a negative binomial zero-inflated regression model. One can see that between 20 and 80 years the total disease count can be fitted very well by this model. On a logarithmic scale, the average

1 R is a free software environment for statistical computing (see http://www.r-project.org/).



(a) All patients



(b) Diabetics

Figure 3.2: Total disease counts in general practice in the Netherlands; shown are the average number of chronic diseases by age on a logarithmic scale. The black line represents the average observed disease count for the whole population, red represents females, and blue males. Dashed lines represents estimation by negative binomial zero-inflated regression using the 20-80 year interval. The estimation for all patients is plotted in the results obtained for diabetics (b), to show that their counts are systematically higher.

	All Patients		Diabetics	
Model	AIC	BIC	AIC	BIC
Poisson	955959	955990	84790	84813
Poisson zero-inflated	853709	853750	82495	82526
Negative binomial	827401	827443	78486	78509
Negative binomial zero-inflated	823500	823553	78406	78445
Geometric	827936	827967	82483	82498
Geometric binomial zero-inflated	826898	826940	82487	82518

Table 3.2: Comparison of count models.

disease count has a steady pace: from approximately 1 at 40 years to approximately 5 at 80 years. However, for children the observed disease count is much higher than expected by this model, which is in line with observations of Van den Akker et al. [207]. For the elderly above 80 years, the observed counts are lower than expected. This is partly due to the fact that the dataset is in some sense censored, i.e., patients over 80 years in the population observed here are 'survivors'. For diabetics the starting age at which the model starts to explain the data better is approximately 40 years. Note that, the absolute values of the disease counts depend on the total number of diseases used in the analysis.

To detect possible interactions between diseases that affect the total disease counts we need to model the observed counts more precisely. Assume that a chronic disease D is modelled by the Markov chain of Figure 3.3, and we assume that $p_t = pe^{Ct}$ (for $C = 0$ we obtain the time-homogeneous model). Then, the expectation of D at time $T = t$ is given by:

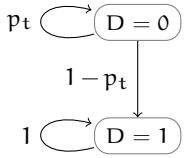
$$E[D \mid T = t] = 1 - p^t e^{(Ct(t-1)/2)} \tag{30}$$

Furthermore, suppose that the ages of the observed patients are *uniformly* distributed over the interval $[0, M]$, with M being the maximal possible age, i.e., by defining T as a uniform distribution over the interval $[0, M]$, then it can be shown that the expectation of D over the whole population is equal to:

$$E[D \mid M = m] = 1 - \frac{1}{m} \sum_{t=1}^{m} p^t e^{(Ct(t-1)/2)} \tag{31}$$

In case of time-homogeneity, i.e., $C = 0$, the summation in Equation 31 corresponds to the geometric series, whose sum equals $(1 - p^m)/(1 - p)$. In that case, accumulating over multiple diseases D_j, with $j = 1, \ldots, K$, the expectation of the total disease count becomes:

$$E[D_{tot} \mid M = m] = K - \frac{1}{m} \sum_{j=1}^{K} (1 - p_j^m)/(1 - p_j) \tag{32}$$



1) age-independent model: $p_t = p$

2) age-dependent model: $p_t = p e^{C_1 + Age(t)/C_2}$

3) history-dependent model: $p_t = p e^{C_1 + His(t)/C_2}$

with $His(t)$ the number of diseases present at $t - 1$

Figure 3.3: Discrete-time Markov chain model of a chronic disease D, with possible disease status 0 (absent) or 1 (present). For each time slice there is chance with probability p_t of staying healthy, and with probability $1 - p_t$ of becoming ill. Once ill, this sustains in the next time slices (probability 1). We speak of a time-homogeneous model (e.g., a) in case p_t is time-invariant, i.e., $p_t = p$, and a time-heterogeneous model (e.g., b and c) otherwise.

When the maximal age goes to infinity the expected value obviously becomes K, i.e., all diseases are present. As this is unrealistic, let us take a close look at the range of ages between fifty and hundred. Suppose we are counting diseases, all with an annual incidence of 0.1%, i.e., $p_j = 0.001$. In case the probabilities p_t are time-heterogeneous – see models 2 and 3 in Figure 3.3 – we choose the model parameters in such a way that the annual incidence is gradually increasing to 0.25% at the age of hundred. The total disease counts by age in percentage of the number of diseases counted are then shown in Figure 3.4.

One can see that the two different time-heterogeneous models are close to each other. There is a slight difference in shape observable; however, if one would try to fit real-world counts, each model will probably not outperform the other. Moreover, when using observational data, disease counts are already biased by disease interactions, and the true annual incidences are therefore hard to estimate. For example, we could have chosen the individual annual incidences for the time-homogeneous model in such a way that it resembles close to one of the time-heterogeneous models.

In summary, because of the exponential nature of disease counts, all regression models with a log or logit link will show a good fit. Therefore, such count models are useful to analyse the total burden of multimorbidity for specific explanatory variables, e.g., age. However, to gain more insight into disease interactions one needs other metrics.

3.6 PAIRWISE ASSOCIATIONS BETWEEN CHRONIC DISEASES

There are several ways to express the association between pairs of diseases. The relative risk is the preferred measure of association in clinical epidemiology and odds ratios are used as an approximation in case-control designs. The popularity of the odds ratio is principally due to the ease of its calculation (being simply the cross-product from a two-by-two table) and to the fact that it provides a good estimate of the relative risk, although when disorders are more prevalent its value becomes progressively larger than that of the risk ratio. For example, narcolepsy, a neurologic disorder, shows associations with many psychiatric disorders, something which has been recently expressed in terms of odds ratios [41, 42, 43]. Odds ratios have also frequently been used in multimorbidity research, see e.g., [177, 68, 19, 215].

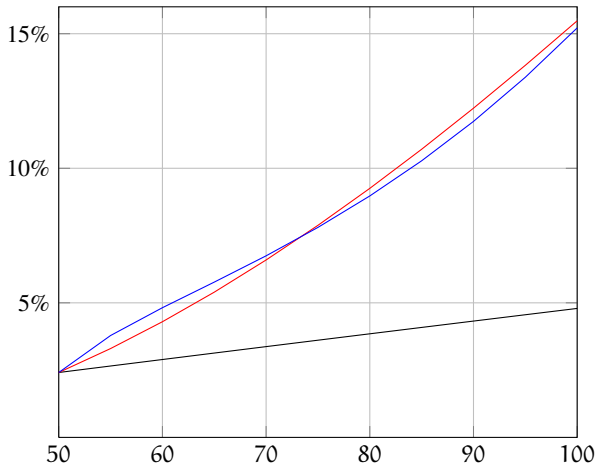


Figure 3.4: Average disease counts, in percentage of the total number of diseases being counted, of the discrete-time Markov chain chronic disease models described in Figure 3.3 for ages between 50 and 100. The black, red, and blue lines correspond with the time-homogeneous, age-dependent, and history-dependent models respectively. For the time-heterogeneous models the parameters C_1 and C_2 are chosen in such a way that they reach the same disease count at age of hundred to make a comparison between the shapes possible.

Odds and risk ratios estimate the overall strength of association between disorders but fail to separate cluster from coincidental comorbidity. The multimorbidity coefficient tries to correct for this phenomenon in as far as it is attributable to coincidental (expected) co-occurrence of disorders. Consider Table 3.3 that is assumed to reflect the number of observations for two diseases, disorder D_1 and D_2. Then the *odds ratio* (OR for short) is defined by:

$$\mathrm{OR} = \frac{ad}{bc}$$

whereas the *multimorbidity coefficient* (MC, discussed in [13]) is defined by:

$$\mathrm{MC} = \frac{a/N}{[(a+c)/N][(a+b)/N]} = \frac{aN}{(a+b)(a+c)}$$

The relative risk (RR) is not symmetric and can be defined for both diseases:

$$\mathrm{RR}_1 = \frac{a/(a+c)}{b/(b+d)} = \frac{a(b+d)}{b(a+c)} \qquad \mathrm{RR}_2 = \frac{a/(a+b)}{c/(c+d)} = \frac{a(c+d)}{c(a+b)}$$

The MC favours pairs of low prevalences, and to lower this tendency a pseudo-count of one can be added to the numerator and denominator of the MC, which was done in the work of Roque et al. [174].

The MC is just an example of many multimorbidity coefficients, and it coincidences with the relative risk for disease pairs (RR_{12}) as defined in the *network medicine* frame-

		Disorder D_1		
		present	absent	totals
Disorder D_2	present	a	b	a+b
	absent	c	d	c+d
	totals	a+c	b+d	a+b+c+d=N

Table 3.3: two-by-two table reflecting the number of observations for two diseases

work, which was recently introduced by Barabási et al. [10]. As an alternative approach they used also the ϕ-correlation coefficient.

$$\phi_{12} = \frac{aN - (a+b)(a+c)}{\sqrt{(a+b)(a+c)(N-(a+b))(N-(a+c))}}$$

From the literature it is known that the odds ratios and relative risks tend to overestimate the association between two variables. In particular, in large samples a statistically significant association may be easily found, although the level of comorbidity is not of clinical importance. Various concordance statistics have been proposed, and just recently a comparison, applied to comorbid diseases, was performed between the Kappa statistic, Somers' D, Kendall's Tau-b, the Gamma statistic, and the adjusted Rand index. It was concluded that the asymmetric Somers' D and Kendall's Tau-b statistics have the highest power to detect non-random comorbidity [139]. These two statistics are calculated in terms of concordant pairs $P = ad$, discordant pairs $Q = bc$, and the tied pairs $T_{row} = (ab + cd)$ and $T_{col} = (ac + bd)$:

$$\text{Somers' D} = \frac{P-Q}{\min(W_{row}, W_{col})} \qquad \text{Kendall's Tau-b} = \frac{P-Q}{\sqrt{W_{row}W_{col}}}$$

with $W_{row} = P + Q + T_{row}$ and $W_{col} = P + Q + T_{col}$. In comparison, the Gamma statistic ignores the tied pairs and is defined as $(P-Q)/(P+Q)$.

Alternatively, odds ratios can also be corrected for other patient related variables by using logistic regression. Marengoni et al. [124] used this methodology to correct the odds for age, sex, education, and other diseases. Nuyen et al. [142] used regression methods to analyse the effects of somatic and psychiatric co-morbidity on depression. However, to the best of our knowledge, there is no such methodology for the multimorbidity coefficient yet.

When looking more closely to the cumulative incidences by age of chronic diseases, they resemble a sigmoid curve. From growth analysis in biostatistics there are some widely used distributions available to model such cumulative incidences. The most common model is the logistic model. This curve is defined by three parameters: the maximal probability A that can be reached when the age $\to \infty$, a maximum slope μ, and the parameter λ representing in some sense the onset. The cumulative incidence of any set D of co-occuring diseases, and thus also a single disease, is then modelled as:

$$P(D = d \mid Age \leqslant t) = \frac{A}{1 + \exp(\frac{4\mu}{A}(\lambda - t) + 2)}$$

with $\mathrm{D} = \mathrm{d}$ meaning that for all $\mathrm{D_i} \in \mathrm{D}$ it holds that $\mathrm{D_i} = \mathtt{True}$. Alternative sigmoid curves are the ones defined by Gompertz and Richards [168].

To illustrate the logistic model, we used it to estimate the prevalences and cumulative incidences of hypertension, diabetes mellitus, and their co-occurrence. Figure 3.5 shows that the logistic models are suitable to model the incidences, in particular the cumulative incidence. Obtaining the expected curve for the co-occurrence of hypertension and diabetes mellitus, by multiplying the individual curves – assuming no interaction between diseases – shows that the observed curve is approximately twice as high, see Figure 3.6.

The cumulative incidence of hypertension or diabetes mellitus in the presence of the other, shows that both estimated curves are higher than in the overall population. However, where hypertensive patients show a more gradual onset of diabetes mellitus, the onset of hypertension in diabetics has a very steep slope between 40 and 60 years, see also Figure 3.6.

3.7 CLUSTERING CHRONIC DISEASES

Clustering techniques are in particular popular in psychiatry; they are often used to link different psychiatric phenotypes to clinical measurements. For example, the phenotypes of autism, attention deficit hyperactivity disorder, and borderline disorder are sometimes hard to distinguish within patients. Especially, when combinations of such disorders occur, clustering techniques provide more insight into the distribution of different combinations of these phenotypes among patients [14, 136, 210, 212, 213].

There are several methods to cluster variables or individuals in a patient dataset. One of them, exploratory factor analysis (EFA), is used to uncover the underlying structure of a relatively large set of variables by the use of latent variables. Its usage to examine co- and multimorbidity within somatic diseases remains subtle. Only recently, EFAs were used as an approach to identify clusters of chronic diseases that share an underlying (hidden) factor [147, 179, 90, 81, 92].

In a confirmatory factor analysis (CFA) one hypothesises beforehand the number of factors in the model, and usually the researcher will also posit expectations about which variables will load on which factors. Johnson and Wolinsky [90] use CFA to explore an interesting model of multiple diseases, disability, functional limitations, and perceived health amongst elderly.

Similar to factor analysis, latent class analysis (LCA) aims at finding a reduced set of dimensions that explains the relations between the variables. Unlike factor analysis, LCA assumes that the latent variable is categorical, and indicators can be nominal. Latent class analysis was used by Schüz et al [182]. Their analysis shows how multimorbidity amongst elderly can be divided into four different profiles.

The grade of membership method, introduced by Manton and Woodbury [123], is used less frequently in multimorbidity research. Portrait et al. [161] used it to measure multimorbidity in the longitudinal ageing study of Amsterdam. In this research, the method tries to link multiple observations to multiple health dimensions.

Here we used LCA to cluster chronic diseases. Using the set of chronic diseases defined in Table 3.1 and the LINH dataset we calculated the decision metrics for solutions up till nine classes, see Table 3.4. It turns out that a solution with eight classes has the

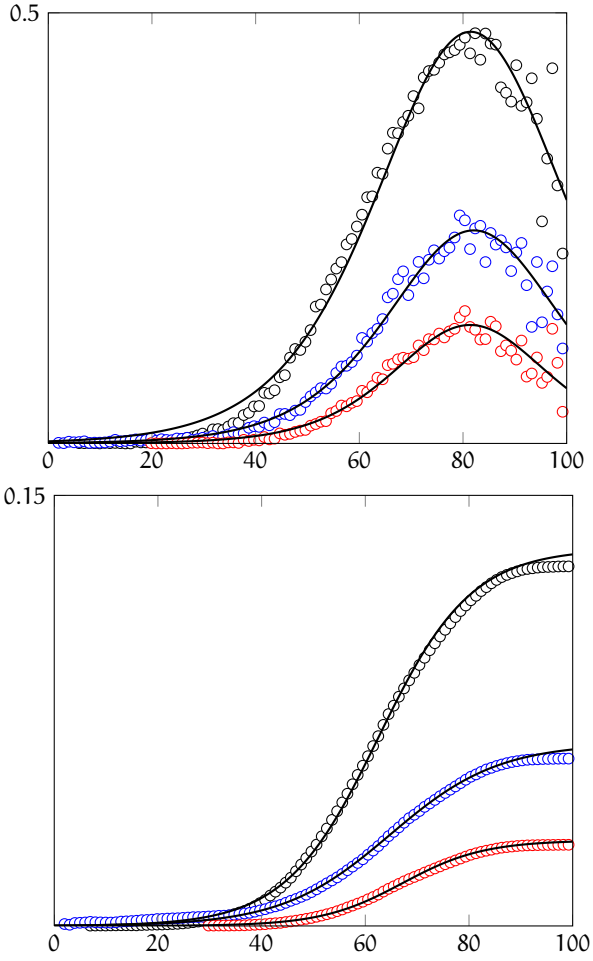


Figure 3.5: Prevalence (top) and cumulative incidence (bottom) by age; for hypertension (black), diabetes mellitus (blue), and their co-occurrence (red); in the Netherlands. Solid lines represent the logistic curves obtained by non-linear least squares estimation using patient data from 90 general practices covering 273,395 patients.

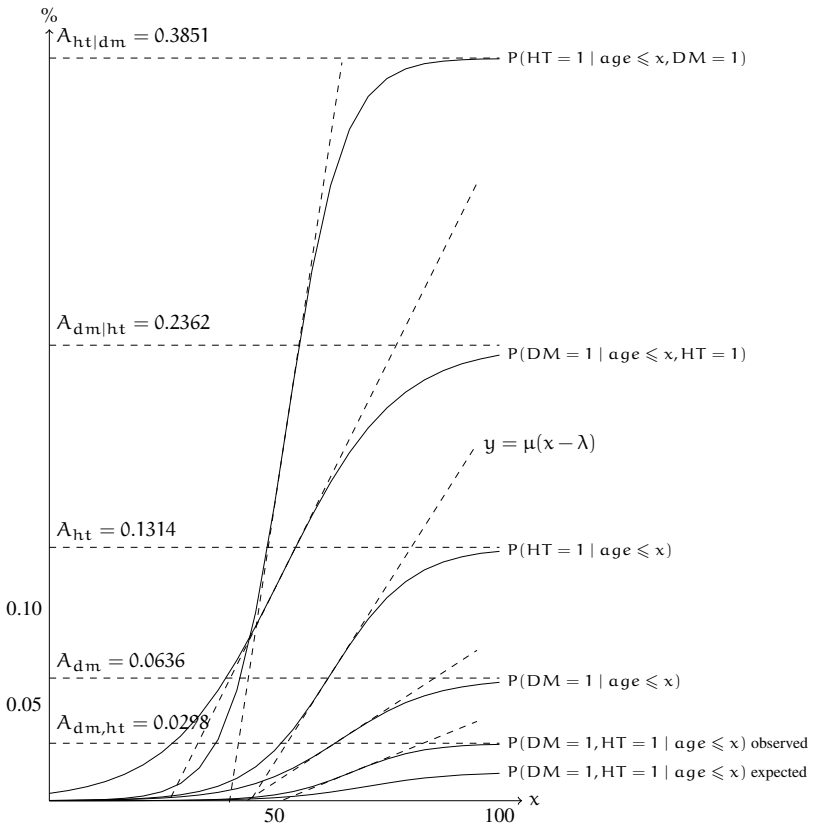


Figure 3.6: Non-linear least square estimations of the cumulative incidences of hypertension, diabetes mellitus, and their co-occurrence. Parameters of the model (other than A) are: $\mu_{ht} = 0.003675$, $\mu_{dm} = 0.001522$, $\mu_{dm,ht} = 0.000952$, $\lambda_{ht} = 44.66$, $\lambda_{dm} = 43.89$, and $\lambda_{dm,ht} = 51.51$.

Number of Classes	AIC	BIC	SSA-BIC	VLMR	p-value LMR-adjusted	p-value
2	1317878	1320570	1319671	0.0000	0.0000	
3	1298402	1302445	1301094	0.0000	0.0000	
4	1292449	1297843	1296041	0.0000	0.0000	
5	1288273	1295017	1292764	0.0000	0.0000	
6	1285151	1293247	1290542	0.0200	0.0202	
7	1283816	1293262	1290106	0.0927	0.0933	
8	**1282275**	**1293072**	**1289466**	0.0000	0.0000	
9	1289534	1293593	1289534	0.0212	0.0214	

Table 3.4: Decision parameters of the LCA results obtained for the LINH data.

lowest values for AIC, BIC, and SSA-BIC. The VLMR and LMR-adjusted test show that the solution is significantly better than the one with seven classes.

Thus, a model in which chronic diseases are clustered in eight groups offers the best explanation for the different profiles of co-occurring diseases one observes in general practice in the Netherlands. Table 3.5 shows all the diseases that had a relatively high prevalence within at least one these groups. Elements of the metabolic syndrome, i.e., hypertension, diabetes mellitus, lipid disorder, and obesity are present in four groups. One can distinguish a group with isolated metabolic diseases, a group with metabolic diseases co-occurring with gastric and musculoskeletal diseases, a group with metabolic diseases co-occurring with cardiovascular diseases, and a group with metabolic diseases co-occurring with cerebrovascular diseases. Then, there are three other groups that cover particular musculoskeletal syndromes, psychiatric disorders, and respirator and dermatologic diseases. The remaining group involves patients having only one or two chronic diseases, which therefore do not contribute to any profile.

3.8 MODELLING EFFECTS OF CHRONIC DISEASES

In case of longitudinal data, one can make predictions about the effect over time, e.g., mortality or quality of life, given a specific set of conditions at baseline. In a regression model, the relation between time and other variables, can be expressed using a dependence on the temporal variable itself and on the product of this variable and all other explanatory variables. The latter represents the interaction between explanations and time, and this model is known as the ‘repeated measures model’. However, in multimorbidity research, Cox regression, also called proportional hazard modelling – the most popular survival analysis model – is mostly used. It is a statistical technique that determines the relationship between survival and several independent exploratory variables. It is useful for modelling the time to a specific event based upon the value of a given covariate.

Disease	Organic System	Class i							
		1	2	3	4	5	6	7	8
% Patients		8.3	2.9	2.6	3.9	9.7	7.5	7.4	57.7
Average Age		64	65	75	75	55	44	26	34
Hypertension	Metabolic	✓	✓	✓	✓				
Diabetes Mellitus	Metabolic	✓	✓	✓	✓				
Lipid Disorder	Metabolic	✓	✓	✓	✓				
Obesity	Metabolic		✓						
Hypothyroidism	Metabolic		✓						
Gout	Metabolic		✓						
Reflux	Gastric		✓						
Irritated Bowel Syndrome	Gastric		✓						
Vertigo	Vestibular		✓						
Spondylosis	Musculoskeletal		✓						
Osteoarthritis	Musculoskeletal		✓						
Shoulder Syndrome	Musculoskeletal		✓						
Tendinitis	Musculoskeletal		✓						
Osteoporosis	Musculoskeletal		✓						
Carpal Tunnel Syndrome	Musculoskeletal		✓						
Spinal Disc Herniation	Musculoskeletal		✓	✓		✓			
NOS	Musculoskeletal		✓	✓		✓			
Varicosis	Cardiovascular		✓						
Angina Pectoris	Cardiovascular			✓					
Myocardial Infarction	Cardiovascular			✓					
Coronary Sclerosis	Cardiovascular			✓					
Atherosclerosis	Cardiovascular			✓					
Atrial Fibrillation	Cardiovascular			✓					
Heart Failure	Cardiovascular			✓	✓				
TIA	Cerebrovascular				✓				
CVA	Cerebrovascular				✓				
Dementia	Cerebrovascular				✓				
Renal Insufficiency	Urologic			✓					
Benign Prostate Hypertrophy	Urologic			✓					
Cataract	Eye			✓					
Neuropathy	Neurologic		✓						
Migraine	Neurologic						✓		
Anxiety	Psychiatric						✓		
Depression	Psychiatric						✓		
Neurasthenia	Psychiatric						✓		
COPD	Respirator			✓					
Tonsillar Hypertrophy	Respirator							✓	
Asthma	Respirator							✓	
Seborrhoeic Eczema	Dermatologic							✓	
Atopic Eczema	Dermatologic							✓	
Acne	Dermatologic							✓	
NOS	Dermatologic		✓	✓					

Table 3.5: Chronic diseases having a clinical relevance in one the classes belonging to the eight-class LCA solution in Table 3.4.

Weight	Conditions
1	myocardial infarct, congestive heart failure, peripheral vascular disease, cerebrovascular disease, dementia, chronic pulmonary disease, connective tissue disease, ulcer disease, mild liver disease, diabetes
2	hemiplegia, moderate or severe renal disease, diabetes with end organ damage, any tumour, leukaemia, lymphoma
3	moderate or severe liver disease
6	metastatic solid tumour, AIDS

Table 3.6: Weighted indices used in the Charlton comorbidity index.

For example, Glynn et al. [61] used Cox's regression to evaluate the effect of the presence of chronic kidney disease and diabetes on mortality and morbidity among patients with established cardiovascular disease. Lin et al. [112] used proportional hazard models to evaluate depression, cardiovascular disease, and increased mortality in diabetes. Putter et al. [163] describe more advanced survival-analysis methods that handles competing risks and multi-states models.

A special case of a Cox regression is the CCI, which is a weighted index that predicts the 10-year survival of patients with multiple disorders. The relationship between potential prognostic diseases and survival was assessed using Cox regression method. The resulting index-methodology works as follows. First, a composite score is determined based on age and the presence of specific comorbid disorders. In this score each decade of age over 40 adds one point. Each disorder from a predefined comorbidity list adds another specific amount of points, see Table 3.6. Secondly, the score is then used to predict the 10-year survival. It is assumed that the overall 10-year survival in an average low-risk population is 98.3%. The predicted 10-year survival probability P_{10} is then calculated as:

$$P_{10} = 0.983^{e^{0.9 \cdot score}}$$

For example, a patient aged 60 years with congestive heart failure and diabetes with end organ damage, has a score $2 + 1 + 2 = 5$. The prediction is than calculated as: $e^{0.9score} = e^{4.5} = 90$ and $0.983^{90} = 0.213$, implying a predicted 10-year survival probability of approximately 21%.

3.9 DISEASE NETWORKS

Recently, networks of associated diseases – referred to as *network medicine* – were proposed as a solution to discover different phenotypes of multimorbidity [10]. In such networks, diseases are represented as vertices, and their associations are represented by undirected edges. In the networks of [10] edges were drawn between each pair of diseases with a significant association based on the multimorbidity coefficient. How-

Behcets syndrome, Reiters syndrome, patella chondromalacia, Pagets (bone) disease, muscular dystrophy, lupus erythematosus, osteitis, osteitis deformans, polymyositis, progressive system sclerosis, repetitive strain injury, diffuse or localised scleroderma.

Table 3.7: Chronic musculoskeletal diseases which are grouped together as 'not otherwise specified' (NOS) in the definition of chronic diseases in [144].

ever, this approach does not correct for confounding effects; a significant association between two diseases can still be caused due to a third disease having an effect on both the others. Learning a structure in such a way that these effects are abandoned from the network, is exactly what is happening when learning the structure of a Markov network.

In case a *direction* is given to all the edges, this is called a Bayesian network. To illustrate the concept of a disease network we took all the diseases from the latent classes present in Table 3.5 and used bootstrapped structure learning of Bayesian networks to build the network. In Chapter 5 we will elaborate more on this technique, for now we only show what it is capable of and the result is presented in Figure 3.7. One can see that most of the edges in Figure 3.7 are between disease vertices that are in the same cluster as determined by the LCA. However, now both the inter-cluster and the intra-cluster interactions between chronic diseases are visualised.

The number of edges in Figure 3.7 is approximately 10% of the maximum number of edges possible for this network, i.e., when all vertices are directly connected which each other. In comparison to the undirected disease network, where edged are determined by pairwise assocations, the number of edges is significantly lower. For example, taking all pairwise associations with a multimorbidity coefficient > 2 results in an undirected network with approximately 60% of the maximum number of edges possible. To obtain the same density of edges, as in Figure 3.7, one should take associations with a multimorbidity coefficient approximately > 7. The same conclusion holds for the odds ratio.

Furthermore, it turned out that the elements of the metabolic syndrome, i.e., hypertension, lipid disorders, and diabetes mellitus, and musculoskeletal disorders NOS are the major confounders of all other chronic disorders. From the metabolic syndrome this was already known. For the musculoskeletal disorders NOS, which are listed in Table 3.7, this was less obvious. It might be that it represents a kind of latent pathophysiology, e.g., an autoimmune disease or genetic predisposition of unknown origin. Something which is comparable with hypertension, of which the aetiology is also not entirely known. In fact, in 95% of the patients with hypertension, the cause is not being identified [25].

Network models can also be useful to detect conflicts within clinical guidelines due to interactions between diseases. For example, the clinical guidelines of hypertension and dementia share common risk factors, symptoms and signs, suggesting there are pathological pathways contributing to both diseases. In treatment there can be many interactions between disorders and drugs, such as positive and negative additive synergies, and drug antagonism when the drugs cancel each other effects. To illustrate this,

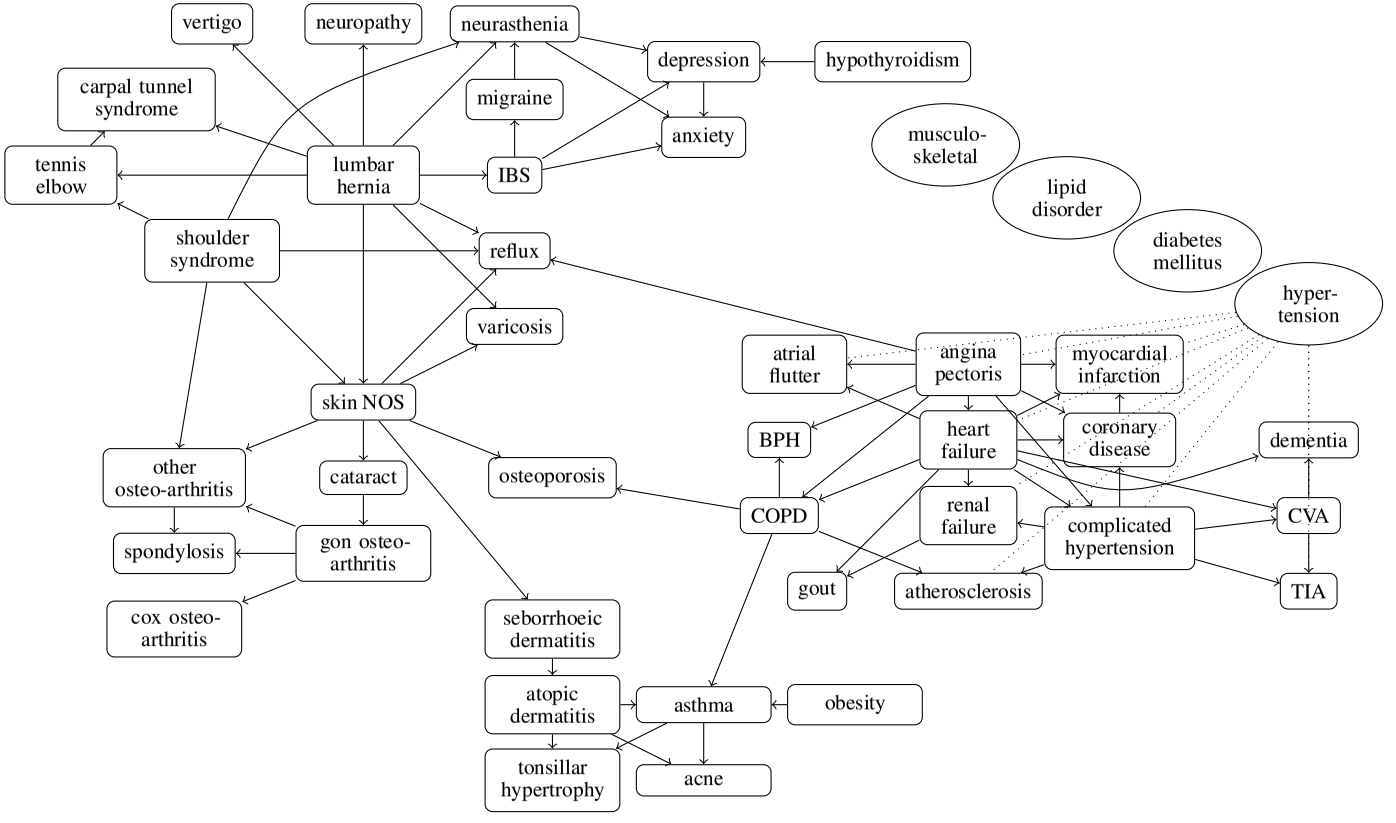


Figure 3.7: Bayesian network of the chronic diseases in Table 3.5. Musculoskeletal diseases not otherwise specified, lipid disorders, diabetes mellitus, and hypertension have edges to most of the other diseases of all clusters; for readability they are not shown.

both guidelines were analysed with the aim of constructing a model that could be used as a start for detecting interactions between hypertension and dementia. The result of such an analysis is shown in Figure 3.8.

From the picture one can see that both diabetes mellitus and cardiovascular diseases have their influence on the pathophysiologic processes of dementia. Not only by their effect on vascular pathologies, but also by their effect on cerebral perfusion. The clinical meaning of this is that, treating one disease affects the other. For example, blood pressure regulation by antihypertensive drugs can affect the onset of dementia both in a positive and in a negative way. Indeed, high blood pressure may accelerate cerebral white matter lesions, but white matter lesions have also been found to be facilitated by excessive fall in blood pressure, including orthostatic dysregulation and postprandial hypotension [135]. Moreover, just recently, it was shown that one of the blood pressure regulation mechanisms – more precise: the baro-reflex – is compromised in patients with Alzheimer. Pharmacotherapeutic agents (cholinesterase inhibitors) for the treatment of dementia, also partly restored this blood pressure regulation mechanism [101].

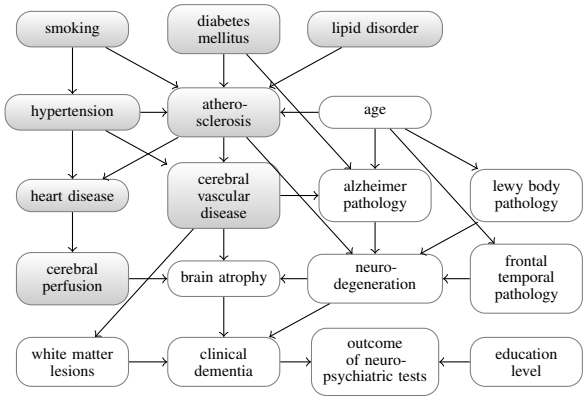


Figure 3.8: Simplified network model derived from cardiovascular and dementia clinical guidelines.

3.10 DISCUSSION

In this chapter, we provided information about the storage of patient data in general practice into electronic health records. A part of this large collection of patient data – called the LINH data – is used for analysis throughout this thesis. Secondly, we reviewed and discussed the current methodologies that are frequently used in multimorbidity research, and illustrated some of them by applying them to the LINH data.

Total disease counts are useful to provide insight into how they evolve when people age. Except for childhood and ages above eighty, total disease counts for the LINH data were proportional with age on a logarithmic scale. However, from the total figures one cannot determine whether this exponentional growth is due to disease interactions or due to the effect of ageing itself.

Pairwise disease associations provide more insight into the interaction between diseases. Although there are several metrics in this area, in multimorbidity research the multimorbidity coefficient (MC) is preferred. Odds ratios can be estimated using logistic regression, which makes them capable of correcting for age or other explantory variables. However, the MC cannot be corrected in the same way. At a specific age, the MC is reflected by the individual cumulative incidences (CI) and the CI of its co-occurrence. Modelling the CI by a sigmoid curve we showed that the product of the individual curves – reflecting the expected CI of the co-occurrence – can be easily compared with the observed CI. Here we showed that the observed CI of diabetes mellitus and hypertension together was significantly higher than expected for ages above fifty.

The next step in multimorbidity research is very often the detection of groups of diseases, assuming there are latent variables that are responsible for the observed associations. Here we applied a latent class analysis on the LINH data, showing that chronic diseases can be divided into eight groups. Elements of the metabolic syndrome are present in four of these groups, stressing out their impact on multimorbidity figures.

Using the pairwise associations one can build a network in which diseases are connected when the observed association reaches a pre-defined significance. However, in this way conditional independence is not always incorporated. From probabilistic graphical modelling they are ways to learn the conditional independences. Here we applied Bayesian network structure learning – something we will elaborate more on in Chapter 5 – on the set of chronic diseases that were most discriminating in the latent class analysis. Hypertension, diabetes mellitus, lipid disorders, and musculoskeletal disorders turned out to be responsible for many pairwise disease associations, making the resulting network more comprehensible than a network of pairwise associations.

In Chapter 6 we will show how Bayesian networks can be made suitable for dealing with the fact that patient data is biased due to practice related effects. In this chapter, we also discussed some temporal models used in multimorbidity. In Chapter 7 we will elaborate more on this by exploring the use of temporal Bayesian networks.

4

PROBABILISTIC CAUSAL MODELS OF MULTIMORBIDITY CONCEPTS

ABSTRACT

Multimorbidity, i.e., the presence of multiple diseases in one person, is a significant healthcare problem for western societies: diagnosis, prognosis, and treatment in the presence of multiple diseases can be complex due to the various interactions between diseases. A literature review revealed that there are a variety of definitions that describe different concepts with respect to multimorbidity, both for the cause of multimorbidity as well as for the implications of multimorbidity. To develop computerised decision-support systems that are able to provide personalised patient care, and that would be suitable for replacement of current guideline documents, multimorbidity aspects need to be captured rigorously in a formal language. In this chapter, we employ causal Bayesian networks to propose a novel framework that can be used to model a spectrum of aspects of multimorbidity. We conclude that this framework offers a foundation for modelling interactions between multiple diseases.

4.1 INTRODUCTION

As discussed in the previous Chapter, there are various ways to measure *multimorbidity*. The use of indices, which has its origin in the analysis of *comorbidity*, see also Section 1.4, already shows a large variety in the point of view of researchers. Recently, a systematic literature research [39] emphasised this heterogeneity in multimorbidity indices. Although these indices show us the size, impact, and growth of the multimorbidity burden, they do not give much insight into the underlying causal relationships between different chronic diseases that occur simultaneously in patients.

The need for integral, optimal management of a patient with multiple diseases, and the need to do so using decision-support technology, implies the need for an integral research methodology for multiple diseases. It is unlikely that such methodologies will be based upon traditionally statistical methods, such as logistic regression, as this focuses on the predictive power of specific variables for the presence or absence of one particular disease [218]. In this chapter, we will argue that probabilistic graphical models, such as Bayesian networks [149], provide a good starting point for modelling interactions between multiple diseases. The edges of a graphical model represent statistical

relationships between variables, which generalises to multiple diseases in a natural way.

Some examples exist in clinical research that model specific diseases within a multimorbidity setting [162, 184, 217]. However, to provide a more generic framework, we need precise probabilistic definitions of the existing concepts in multimorbidity. Firstly, we summarise existing classifications and terminologies used in definitions used in multimorbidity research and point out their similarities and differences. Secondly, we will provide a rigorous probabilistic framework of multimorbidity concepts, using causal Bayesian networks, that fits these classifications and terminologies. Existing definitions of multimorbidity aspects are analysed on the basis of this framework.

4.2 BACKGROUND

4.2.1 *Comorbidity and multimorbidity*

In this chapter, the principal focus is on multimorbidity, but since this is closely related to comorbidity, we also study the concepts related to comorbidity. Recall that comorbidity was defined in relationship to a specific index condition by Feinstein [50]. Details about multimorbidity and comorbidity can be found in Chapter 3. As argued there, much of the medical research relies on regression models which are applied to a single disease, and, thus, ignore the complexity of multimorbidity. Prevalence of multimorbidity has been studied in family practices [207, 53], sometimes with clustering of specific diseases [124], or a factor analysis to reveal patterns of co-occurrence of diseases [179]. These methods show that cardiovascular diseases often co-occur with metabolic diseases, and psychiatric diseases often co-occur with neurologic and somatic diseases causing chronic pain and disability.

These results illustrate the impact and complexity of multimorbidity, but give little insight into interactions between diseases. A systematic review on ageing with multimorbidity [125] identified twelve cross-sectional studies on multimorbidity, four on incidence and risk factors for multimorbidity, twenty-two on consequences of multimorbidity, nine on function status, six on quality of life, eight on healthcare utilization and six on models and quality of care. One of the major conclusions is that little is known about causality in multimorbidity.

Recently Valderas et al. [204] summarised several conceptual problems. Differentiating the nature of conditions is critical to the conceptualisation of comorbidity. For example, conditions can be part of a certain syndrome and should perhaps not be classified as having comorbidity. The question of which condition should be designated as the index and which as the comorbid condition is not self-evident and may vary in relation to the research question, the disease that prompted a particular episode of care, or of the speciality of the attending physician. In that respect, the sequence in which comorbidities appear may have important implications for genesis, prognosis, and treatment.

From a patient and physician's point of view, multimorbidity is part of a bigger concept, i.e., the multimorbidity burden, which adds parameters such as polypharmacy, sex, age, frailty and other health-related individual attributes and the patient's complexity (adding non-health-related individual attributes). For polypharmacy, multiple

definitions are utilised in the literature. Basically, a certain minimal number of drugs have to be used, but additional definitions include the minimal time of subscription, regular daily consumption of multiple medications, and the use of high-risk medications and questionable dosing. A literature review on polypharmacy in the elderly [58] stated that selecting appropriate limits for numbers of medications may be counter-productive in populations with multiple comorbidities.

4.2.2 *Causal relations within multimorbidity*

A part of the context of the multimorbidity burden is illustrated by Figure 4.1, which provides an abstract view on the problem. We make the assumption that a disease always corresponds to a particular pathophysiology, in contrast to syndromes. Syndromes represent a certain symptomatology which can be caused by several pathophysiologic processes. Furthermore, within gerontology certain combinations of diseases are defined as a geriatric syndrome, meaning a combined set of specific symptomatology that leads to impaired daily functioning. For example, the combination of polyneuropathy, impaired vision, and the usage of drugs that affect the patient's consciousness (e.g., benzodiazepines), often leads to higher risks of falling, the latter being defined as a geriatric syndrome.

A therapy does not necessarily have to act directly on the underlying pathophysiology of the disease intended to treat. In many cases another physiologic process is used to suppress the symptoms of the disease. For example, within hypertension diuretics use kidney function to lower the blood pressure, although the actual cause of the hypertension within a particular patient may be due to another pathophysiologic process. Another issues is that where a therapy typically acts on a designated (patho)physiology, it may also act on another physiologic processes, causing side effects. Accumulation of side effects due to polypharmacy can have a major impact on the quality of life.

Whereas the single disease model is fairly simple, mutual dependences within the multiple disease model may concern the pathophysiology, symptomatology, therapy, and prognosis. By modelling these interactions explicitly, better decisions can be made for patients who have multiple diseases. Moreover, single disease models often contain a lot of overlap; this redundancy may be avoided by integrating different disease models into a single model. For example, consider a physician facing a patient with a history of multiple chronic diseases, now having a new problem. How could the physician tell if the problem is caused by either a new disease, existing morbidity, a side effect of existing pharmacotherapy, or just a natural phenomena due to ageing or ageing related stress factors? Treating it just as a new problem (whether introduced by existing diseases and treatments or not) is often the most pragmatic solution. However, with each newly introduced treatment the overall personal multimorbidity puzzle becomes more and more complicated.

As an example consider diabetes mellitus (DM), in which two types are distinguished: type I, due to impaired insulin production caused by destruction of β-cells in the pancreas, and type II, due to insulin resistance of peripheral tissue. Both types cause uncontrolled high blood glucose levels. Measuring this feature of the disease, i.e., high blood glucose levels, is the corner-stone of the diagnosis and pharmacological control of DM. The main consequences of sustained high blood glucose levels are

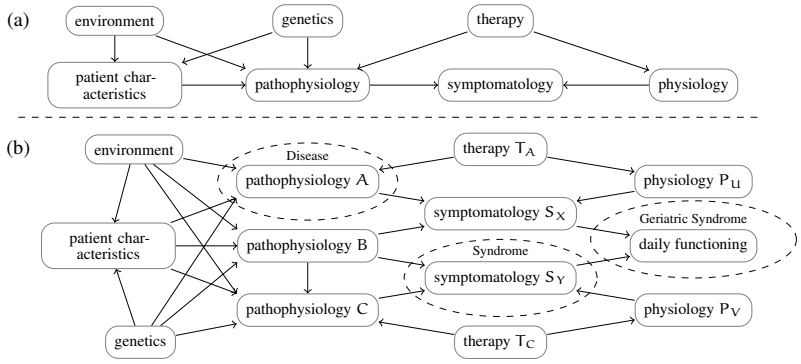


Figure 4.1: Abstract model of a single disease (a) and multiple diseases (b).

neuropathy and blood vessel damage, the latter causing new diseases, e.g., retinopathy, renal failure, heart failure, impaired wound healing etcetera. So, a sustained pathophysiological condition caused by a chronic disease, which can be measured using a specific laboratory test, is often the cause of new diseases. In fact, we can do the same exercise for condition like a sustained high blood pressure (mostly of unknown cause) or high levels of blood lipids (mostly of dietary cause). In general practice, the presence of such secondary diseases strengthens the diagnosis.

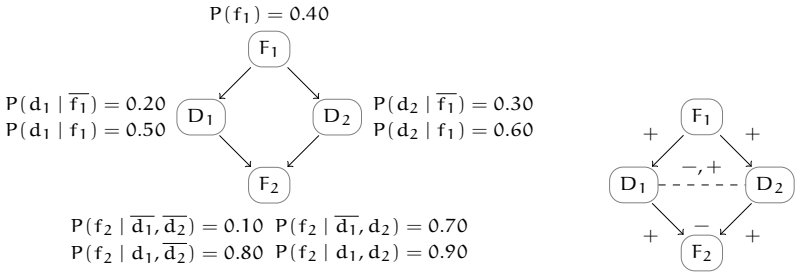
4.3 METHODS

To model and analyse multimorbidity concepts, we will employ Bayesian networks (for details see Chapter 2), which have the ability to model more complex structures involving disease variables than traditional regression models. While regression models can only represent functions with just one dependent variable, e.g., a single disease, Bayesian networks allow for inference about multiple diseases at the same time. Moreover, it has been shown that in complex medical domains, Bayesian network can outperform the predictive power of regression models [105].

Causal Bayesian networks are Bayesian networks where the directed edges in G represent causal influences between variables [151], i.e., an arc between C and E means that C is a cause of E. In these models, we can consider probability distributions after interventions, written as $P(x_1, \ldots, x_n \mid do(x_i))$, e.g., a probability distribution after modifying a certain risk factor. This probability distribution can be computed by:

$$P(x_1, \ldots, \hat{x}_i, \ldots, x_n \mid do(x_i)) = \begin{cases} \prod_{v \in V, v \neq i} P(x_v \mid x_j \text{ for all } j \in \pi(v)) & \text{if } \hat{x}_i = x_i \\ 0 & \text{otherwise} \end{cases}$$

In order to represent qualitative relationships between variables, we use qualitative causal influences and synergies resembling those of qualitative probabilistic networks [221, 76] (see Section 2.3). The semantics of these qualitative signs is slightly different



(a) Causal Bayesian network

(b) Qualitative influences

Figure 4.2: Example domain with two diseases and two features: F_1 causes D_1 and D_2; D_1 and D_2 both cause F_2. On the left a causal Bayesian network with its associated conditional probability distribution. On the right, the same network with its qualitative signs. All the influences between variables are positive, the additive synergy between D_1 and D_2 is negative and the product synergy between D_1 and D_2 is negative if f_2 and positive if $\overline{f_2}$.

in order to express comorbidity and multimorbidity concepts. Concretely, if we have an arc $C \rightarrow E$, then we say C *causally positively influences* E if:

$$P(e \mid do(c)) > P(e \mid do(\overline{c}))$$

Negative causal influences can be defined similarly. We can also consider synergies, e.g., *positive additive synergies* express that the joint causal influence of C and C' is greater than their separate influence on their child E, i.e.,

$$P(e \mid c, c') + P(e \mid \overline{c}, \overline{c}') > P(e \mid c, \overline{c}') + P(e \mid \overline{c}, c')$$

Finally, we define *causal product synergies* that expresses how the value of one variable influences the probability of the values of another variable in view of a third variable. A negative product synergy of C and C' on variable E with value e means that if C is the case, then this renders C' less likely, which can be expressed by:

$$P(e \mid c, c') \cdot P(e \mid \overline{c}, \overline{c}') < P(e \mid c, \overline{c}') \cdot P(e \mid \overline{c}, c')$$

Similarly, negative additive and positive product synergies can be defined. See Figure 4.2, which illustrates a causal Bayesian network and the qualitative influences and synergies derived from the probability distribution.

4.4 RESULTS

We will first list existing concepts with respect to comorbidity and multimorbidity from literature. Then we systematically discuss causal network structures with qualitative signs and provide an formal analysis with respect to these existing concepts.

4.4.1 *Existing comorbidity and multimorbidity concepts*

We searched the literature for possible relationships between comorbid and multimorbid diseases. The result of this search is summarised in Table 4.1. Each of these papers

introduces concepts related to comorbidity and multimorbidity, which are sometimes mutually exclusive, but often overlapping.

The classification made by Kraemer [98], was one of the first, classifying comorbidity into random, clinical (C-comorbidity), familial (F-comorbidity), and epidemiologic comorbidity (E-comorbidity). While clinical and familial comorbidity are defined, the focus of this paper is mostly on the measurement and interpretation of E-comorbidity. An analysis of C-comorbidity and F-comorbidity is not given. The classification proposed by Van den Akker et al. [207], based on the categorisation of Schellevis [180], is a hierarchical classification. Obviously, all instances of comorbidity fulfil the definition of concurrent comorbidity. Some of these comorbidities will occur in numbers greater than expected by chance and hence should be classified as cluster comorbidity. Some of those statistically significant comorbid associations represent known causal relationships and should be defined as causal comorbidity. This classification is taken significantly further in the work by Valderas et al. [204]. In this paper there are three ways that lead to associations between diseases: by direct causation, by associated risk factors, or by heterogeneity in risk factors. In the direct causation model, the pathophysiology of one disease leads to another disease. In the associated risk factor model, risk factors are correlated (e.g., one causes another). Finally, in the heterogenic risk factor model, the risk factors are independent, but influence both diseases. For two given diseases, these models can occur at the same time.

Besides the mechanisms for the co-existence of multiple diseases, differences in implications for multimorbid diseases on clinical care is also relevant. For example, Kaplan et al. [50, 91] distinguished between diagnostic and prognostic comorbidity. In diagnostic comorbidity different diseases can share specific symptomatology, making the diagnosis harder. In prognostic comorbidity, comorbid diseases alter the prognosis of the patient (mostly negative), sometimes as expected (cogent), but sometimes also unexpected (non-cogent). Closely related to diagnostic comorbidity is the work done by Angold et al. [7], who classified comorbidity into homotypic and heterotypic. These terms are typically used in psychiatric comorbidity, where in homotypic comorbidity, diseases belong to the same diagnostic group, e.g. depression and a dysthymic disease, and in heterotypic comorbidity, diseases belong to a different diagnostic group, e.g. depression and a personality disease. Finally, Piette et al. [159] classified multiple diseases into concordant and discordant. Concordant diseases are part of the same pathophysiology, e.g., cardiovascular diseases due to atherosclerosis, or share a specific therapy, e.g., β-blockers are used for both hypertension and cardiac arrhythmias. Discordant diseases do not share a part of the pathophysiology or similar types of management.

4.4.2 *Aetiological probabilistic models of multimorbidity*

From a probabilistic point of view, there are only a few essential differences in the co-occurrence of multiple diseases. The first possibility is that the co-occurrence between diseases is random, which means that the co-occurrence of these diseases is exactly from what can be expected by chance. Adopting the standard probabilistic terminology,

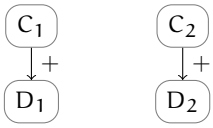


Figure 4.3: Non-aetiological multimorbidity, where diseases D_1 and D_2 are caused by independent factors C_1 and C_2.

we will call this notion *independent multimorbidity*, and can be expressed in probability theory by:

$$P(d_1, d_2) = P(d_1)P(d_2)$$

Assuming the binary random variables D_1 and D_2 represent the occurrence of each disease, it follows that D_1 is independent of D_2, e.g., the absence of d_2 also will not have an influence on d_1 as

$$\begin{aligned} P(d_1 \mid \overline{d_2}) &= \frac{P(d_1, \overline{d_2})}{P(\overline{d_2})} = \frac{P(d_1) - P(d_1, d_2)}{1 - P(d_2)} \\ &= \frac{P(d_1) - P(d_1)P(d_2)}{1 - P(d_2)} = \frac{P(d_1)(1 - P(d_2))}{1 - P(d_2)} = P(d_1) \end{aligned}$$

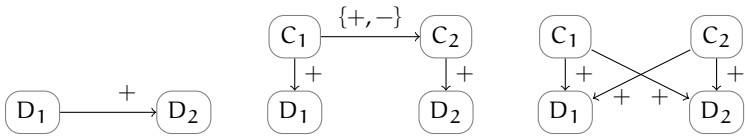
This notion completely coincides with the notion of random co-occurrence [98] and non-aetiological associations [204]. In these models, there is no direct causation between diseases, nor are the causes of the diseases related. See Figure 4.3 for a simple example of non-aetiological multimorbidity. In non-aetiological multimorbidity models, it holds that D_2 is d-separated from D_1 by $\varnothing$. Therefore, it holds that D_1 and D_2 only occur because of independent multimorbidity.

We propose to call the opposite of independent multimorbidity *associative multimorbidity*, which means that there is some relationship between two diseases which cause co-occurrence of diseases to be different from expectation by chance. Formally speaking, in the terminology of Kraemer, this notion is called *epidemiological comorbidity* [98]. However, Kraemer requires that this association should also be epidemiologically measurable (making it independent from F-comorbidity, see below). Typically, one is interested in *positive* associative multimorbidity, i.e., if

$$P(d_1, d_2) > P(d_1)P(d_2).$$

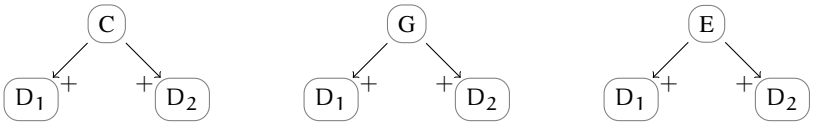
This coincides with what Van den Akker et al. call *cluster comorbidity*, i.e., if d_1 is the index disease, then d_2 is more likely to occur than expected if d_1 and d_2 would have been independent. Negative associative multimorbidity can also occur, e.g., it seems that myopia is protective against diabetic retinopathy [111]. For patients with diseases that are negatively associated, Van den Akker et al. then speak of concurrent comorbidity as this co-occurrence is caused by 'chance' rather than the association.

The division into types of association according to Valderas et al. [204] is listed in Figure 4.4. The direct causation model coincides with the definition of causal comorbidity by Van den Akker et al. [207], whereas the other two (associated risk factor model and heterogenetic risk factor model) are considered cluster comorbidity. From



(a) Direct causation (b) Associated risks (c) Heterogenic risks

Figure 4.4: Example causal network structures of comorbidity relations as defined in by Valderas et al. [204] where diseases are associated. In the associated risk factor model, C_1 and C_2 can be associated by any causal mechanism.



(a) Patient characteristic (b) Genetic (c) Environmental

Figure 4.5: Primary classes of single risk factors inducing multimorbidity, with D_i diseases, C a patient characteristic, G a genetic factor, and E an environmental factor.

a formal point of view, the focus in these models on risk factors, rather than causes of diseases could be considered problematic. For example, the authors write that if "the risk factors for 1 disease are correlated with the risk factor for another disease" (i.e., in the associated risk model) then this makes "the simultaneous occurrence of the diseases more likely". However, this conclusion is only valid if the risk factors are of a causal nature as it is depicted in Figure 4.4. Consider, e.g., diabetes and familial hypercholesterolemia, which both causes elevated LDL cholesterol. While an elevated LDL cholesterol is a (non-causal) risk factor for both diseases, these diseases themselves are not associated.

Besides the way in which the association between diseases is structured, we can also consider which types of causes lead to a positive association between diseases. In Figure 4.1 we introduced an abstract framework with mechanisms that diseases can be related, which directly yields number of possibilities. In general, there are a large number of possible combinations of risk factors. Focusing on a single disease, there are three primary causal risk factors: environmental, genetic and those related to patient characteristics. Example models are given in Figure 4.5. In biomedical research, they play a distinct role: in practice, controlled studies are only performed to study influences of patient characteristics on diseases; environmental factors can only be studied well using epidemiological research; finally, genetic factors are often researched in observational studies using DNA samples, or twin studies.

Each of the single risk factors causes associations between diseases. Nonetheless, as mentioned, Kraemer notes that *familial comorbidity*, i.e., comorbidity which has a genetic cause, is compatible with independent multimorbidity, i.e., familial comorbidity is compatible with the absence of epidemiological comorbidity. However, this is only true if in some families there is a positive association between diseases and in

some other families there is a negative association between the same diseases. Further, the association may not be epidemiologically relevant if the particular gene causing an epidemiological comorbidity has a low prior probability. Consider for example a familial comorbidity between D_1 and D_2 caused by a genetic factor G. If $P(g) \approx 0$ and $P(D_i \mid \overline{g}) \approx P(D_i)$, then:

$$\begin{aligned} P(D_1, D_2) &= \textstyle\sum_G P(D_1 \mid G)P(D_2 \mid G)P(G) \\ &\approx P(D_1 \mid \overline{g})P(D_2 \mid \overline{g})P(\overline{g}) \approx P(D_1 \mid \overline{g})P(D_2 \mid \overline{g}) \approx P(D_1)P(D_2). \end{aligned}$$

Clearly, familial comorbidity might not be epidemiologically measurable if the diseases are almost independent.

4.4.3 *Probabilistic models for reasoning about clinical impact of multimorbidity*

In practice, the impact of multimorbidity might be more relevant than the actual cause of the co-existence of multiple diseases. The literature describes several dimensions through which the interactions between diseases may be relevant, namely if there are diagnostic, prognostic, or therapeutic interactions.

We say that there is a *diagnostic multimorbidity problem* if there are interactions between diseases that complicate the diagnosis of one of these diseases. In essence, making such a diagnosis involves the consideration of multiple diseases that might be the cause of the presented symptomatology within a patient. While in case of single disease management, eventually one disease from this set of diseases is considered to be the one and only cause of the presented symptomatology, in case of diagnostic multimorbidity it might be the case that more than one disease is involved in the presented symptomatology. Formally, given a sign or symptom S, the diagnostic value of S for a given disease D is typically defined by the so-called diagnostic odds ratio, i.e.,

$$\mathrm{DOR}(s \mid d) = \frac{\mathrm{odds}(s \mid d)}{\mathrm{odds}(s \mid \overline{d})}$$

Multimorbidity has an impact on the diagnostic value of S for D_1 in the presence of another disease D_2 if it alters its diagnostic odds-ratio, e.g., negatively, which can be expressed formally by

$$\frac{\mathrm{odds}(s \mid d_1, d_2)}{\mathrm{odds}(s \mid \overline{d_1}, d_2)} < \frac{\mathrm{odds}(s \mid d_1, \overline{d_2})}{\mathrm{odds}(s \mid \overline{d_1}, \overline{d_2})}$$

which can be shown to be equivalent to:

$$P(s \mid d_1, d_2) \cdot P(s \mid \overline{d_1}, \overline{d_2}) < P(s \mid \overline{d_1}, d_2) \cdot P(s \mid d_1, \overline{d_2})$$

i.e., a negative product synergy. Valderas et al. [204] gives an example of such a problem: patients with diabetes mellitus (*dm*) may have altered pain sensation, e.g. angina pectoris (*ap*), thereby interfering with and making it more difficult to diagnose coronary heart disease (*chd*). It thus holds that:

$$\frac{P(ap \mid chd, dm)}{P(ap \mid \overline{chd}, dm)} < \frac{P(ap \mid chd, \overline{dm})}{P(ap \mid \overline{chd}, \overline{dm})}$$

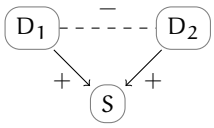
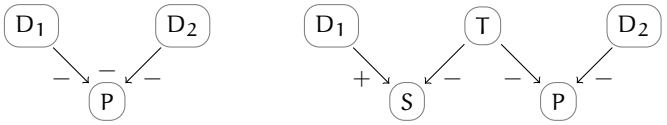


Figure 4.6: Diagnostic multimorbidity problem: D_1 and D_2 positively influence S. A diagnosis for D_1 negatively influences the diagnostic value of s for D_2, and vice versa.



(a) Synergistic prognostic multimorbidity problem

(b) Therapeutic prognostic multimorbidity problem

Figure 4.7: Two types of prognostic multimorbidity. In the synergistic prognostic multimorbidity problem, the negative effects on the prognosis P has an additional negative additive synergy, i.e., the prognosis is worse with multiple diseases compared to the effects of the single disease on the prognosis. In the therapeutic prognostic multimorbidity problem, the therapy given for D_1 – as it negatively influences the symptom S – has a negative impact on the prognosis P for D_2. The effect of T on S is mediated by either the pathophysiology of D_1 or by some other physiological process (cf. Figure 4.1).

expressing that *ap* has less diagnostic value for *chd* in the presence of *dm*. Such diagnostic multimorbidity can be illustrated by a product synergy as shown in Figure 4.6.

A *prognostic multimorbidity problem* occurs if diseases have a (negative) influence on the prognosis of another disease, due to the anticipated effects on therapy and synergetic influence on prognostic factors such as quality adjusted life expectancy (QALYs). This definition is similar to the definition of prognostic comorbidity by Kaplan et al. [91]. For example, diabetes mellitus (DM) and lipid disorders (LD) contribute to biochemical processes that lead to vascular dementia and Alzheimer dementia (AD). Moreover, there is an additive synergy between these two diseases on the prognosis of Alzheimer dementia [100], i.e.,

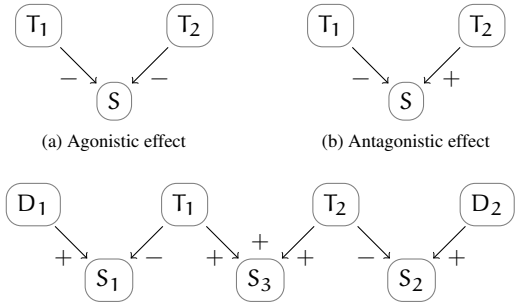
$$\mathrm{P}(\mathrm{p}_{\mathrm{AD}} \mid dm, ld) - \mathrm{P}(\mathrm{p}_{\mathrm{AD}} \mid \overline{dm}, ld) < \mathrm{P}(\mathrm{p}_{\mathrm{AD}} \mid dm, \overline{ld}) - \mathrm{P}(\mathrm{p}_{\mathrm{AD}} \mid \overline{dm}, \overline{ld})$$

Moreover, the prognosis of a disease can be influenced by therapeutic effects given for another disease. For example, using the example of Valderas et al. [204], corticosteroids (*cs*) prescribed for chronic obstructive pulmonary disease in the same patient will have an antagonistic effect on the prognosis of diabetes (p_{DM}), i.e., it is a negative causal influence:

$$\mathrm{P}(\mathrm{p}_{\mathrm{DM}} \mid \mathrm{do}(cs)) < \mathrm{P}(\mathrm{p}_{\mathrm{DM}} \mid \mathrm{do}(\overline{cs}))$$

Examples of such models are illustrated in Figure 4.7.

A *therapeutic interaction problem* occurs when two therapies interact with each other. Agonistic and antagonistic effects can be modelled using qualitative causal influences (cf. Figure 4.8a and 4.8b), although the diseases that are intended to be treated



(a) Agonistic effect

(b) Antagonistic effect

(c) Interaction that induces new symptomatology

Figure 4.8: Therapeutic interactions between independent diseases. In (c) therapy T_1 is used to treat D_1; therapy T_2 is used to treatment D_2; T_1 and T_2 lead to synergetic side-effects (expressed in S_3).

could be independent of each other from a pathophysiologic point of view, i.e., independent multimorbidity. Sometimes, combinations of therapies can even induce new problems. For example, the combination of diuretics (as anti-hypertensive treatment) and NSAID's (as treatment for some independent pain syndrome), can easily lead to dehydration within the elderly. Schematically, this is shown in Figure 4.8c. Hypertension, diuretics and blood pressure, are then represented by D_1, T_1, and S_1 respectively. The same applies to the pain syndrome, represented by D_2, T_2 and S_2. The variable S_3 then represent a side effect of both T_1 and T_2, which prevalence can be even higher than expected due to synergistic effects between T_1 and T_2.

4.5 DISCUSSION

In this chapter we reviewed multimorbidity concepts and proposed a framework using causal Bayesian networks that define many of these concepts in a precise and formal manner. These results show that sometimes concepts are similar from a probabilistic point of view, whereas in other cases, concepts can be sub-categorised into different types of causal networks. The advantage of translating multimorbidity concepts into a causal Bayesian network form is that it supports the explicit modelling of the dependences and independences between the disease variables of concern. This way we obtain a basis for decision-support systems that can be used to manage multimorbidity in patients. While in this chapter we focused on multimorbidity with two diseases, the results presented here generalise to more complex situations, e.g., by also considering qualitative synergies in the presence of more than two causes.

Some concepts from the literature are not explicitly modelled in this framework. Firstly, the definition of clinical comorbidity by Kraemer [98] can be used to designate a variety of comorbid concepts. She states that any disease that has an altering effect on some kind of response to an index disease is of clinical importance. In this definition the response variable can be anything, e.g., age of onset, therapeutic response, prognosis,

which, in a causal network framework, are special cases of prognostic and therapeutic multimorbidity problems. This explains why this concept is not explicitly mentioned.

A similar remark can be made for homotypic and heterotypic multimorbidity. From a diagnostic point of view, homotypic diseases are closer related to each other than heterotypic diseases. This classification has its origin in the aetiology of diseases where the biological substrates are yet unknown to clarify the observed symptomatology, e.g., psychiatric disorders. Very often, latent variable methods, e.g., the ones described in section 2.5.3, are used to explain the observed differences in symptomatology. These models cannot be translated to *causal* Bayesian networks; after all, the latent variables cannot be observed. Still, the technique of using latent variables can be useful for meaningful predictions in case of a general Bayesian network. In case there are many arcs between the latent variables and the observed variables, i.e., in case of homotypic diseases, this, however, may still lead to diagnostic multimorbidity problems.

Finally, the concepts of cogent and non-cogent prognostic comorbidity can be considered concepts beyond causal or probabilistic meaning, as they rely on expectations of the medical researcher.

In Chapter 8, we consider cyclic models, e.g., when diseases contribute to each other's pathophysiology. From a formal point of view, acyclic models can also be used for this purpose by modelling the progression and interaction of diseases over multiple time slices. Cyclic probabilistic graphical models [151] may also be considered as an alternative for providing a succinct representation of such interactions.

In conclusion, the models presented here provide insight into the different aspects of multimorbidity: both in the aetiological relationships between multiple diseases as well as in the impact that multimorbidity has on clinical practice. Aetiological relationships will be further exploited in Chapter 5, by means of critical risk factors, and in Chapter 6, by means of multilevel analysis. The clinical impact of multimorbidity is further shown in Chapter 7, i.e., we analysed the longitudinal impact of commonly observed health conditions, e.g., hypertension, on the prevalence of cardiovascular diseases.

Author	Classification	Definition
Kraemer [98]	random	co-occurrence is completely random
	clinical	the response to a disease depends on the presence of another disease
	epidemiologic	there is a (un)known mechanism that bonds diseases together
	familial	the occurrence of diseases within a family is higher than one would expect purely based on epidemiological chance
Van den Akker et al. [207]	concurrent	the co-occurrence of an index condition with another health condition whether coincidental or not
	cluster	the co-occurrence of an index condition with another health condition at a significantly higher rate than expected by chance
	causal	the causal mechanism underlying the co-occurrence of an index condition and another health condition is known
Valderas et al. [204]	non-aetiological	there is no aetiological association between diseases
	direct causation	one of the disease may cause the other
	associated	the risk factors for diseases are correlated
	heterogeneity	the risk factors are not correlated but affect both diseases
	independence	the presence of the diagnostic features of diseases is due to another distinct disease
Kaplan et al. [50, 91]	diagnostic	an associated disease can simulate symptoms of the index disease
	prognostic	diseases, in relation to an index disease graded according to their anticipated effects on therapy and life expectancy
	- cogent	comorbid ailments expected to impair a patient's survival
	- noncogent	other ailments
Angold et al. [7]	homotypic	diseases within a diagnostic grouping
	heterotypic	diseases from different diagnostic groupings
Piette et al. [159]	concordant	diseases are part of the same pathophysiologic or management profile
	discordant	diseases that are not directly related in either pathogenesis or management and do not share an underlying predisposing factor

Table 4.1: Classification and terminology of comorbidity.

5

FINDING CRITICAL FACTORS IN DISEASE CO-OCCURRENCE

ABSTRACT

To better understand the problem of multimorbidity, we have used Bayesian-network structure-learning methods to discover how risk factors of diseases interact. First, we determined how sensitive structure learning is to the low prevalence of certain diseases. The constraint-based learning methods appeared to be suitable for learning all the significant associations, even though the arc directions were often incorrect. The search-and-score-based algorithms yielded correct arc directions in most cases. However, for these algorithms only the Akaike information criterion seemed to be suitable as a scoring method. Using the Bayesian information criterion, significant clinical associations were penalised away during the learning process. We applied a search-and-score-based algorithm to several sets of chronic diseases using data from general practices. A novel measure for uncovering structural relationships in the co-occurrence of diseases, so called critical factors, is proposed and studied.

5.1 INTRODUCTION

In this chapter we explore the value of Bayesian-network structure learning algorithms in uncovering interactions between diseases. The aim is to develop a new measure that sheds light on the causal risk factors of multimorbidity. The problem of multimorbidity has been described in detail in Chapter 3. Rather than focusing on interactions between diseases of which much is already known, in this chapter we will use the developed methods to detect disease interactions of which much less is known.

A typical example of such a disease type is cancer, i.e., malignant tumours. Although multimorbidity is increasingly attracting attention from oncologists, yet little is known about the interaction between cancers [170]. As cancer is becoming more and more a manageable chronic disease, and because in the ageing Western society more people are at risk for cancer, there is a growing number of patients with multiple malignancies [127]. These multiple cancers affect the survival estimates based on each tumour site, obviously because a primary tumour may have metastasised, but also because there may be multiple primary malignancies [175]. There are quite some risk factors that are implicated in the development of multiple primary malignant tumours, ranging from ageing, environmental and life-style factors, and genetic predisposition. Both for

the development of clinical prediction models and the prevention of multiple malignant tumours, understanding the cause of their co-occurrence is important.

Several conceptual frameworks of multimorbidity have appeared in literature, offering distinct ways to clarify how diseases are related. Recently, we proposed a new framework of multimorbidity, based on Bayesian networks, that can be used as a basis for modelling a spectrum of multimorbidity aspects and that is described in Chapter 4. In this chapter, we build upon this work by developing a new method for the identification of aetiological interactions between diseases from data. In particular, we propose a number of measures that express the interaction between diseases. Furthermore, we identify the *critical factors* that relate the diseases: these factors indicate which mechanisms best explain their co-occurrence.

We evaluate this approach on the most common co-occurrences of malignancies, and show that we can identify the relationships between malignancies and their critical factors. Before doing so, we first need to determine which of the structure learning algorithms is the most suitable for the used dataset. For this purpose, we compare several structure learning algorithms on large datasets in the following section. Thereafter, we turn back to the aetiologic interactions in oncology.

5.2 STRUCTURE LEARNING OF BAYESIAN NETWORKS AND LARGE DATASETS

5.2.1 *Algorithms and data*

There are several methods to learn the network's structure (see the preliminaries in Chapter 2), and during the last two decades, many new structure learning algorithms for Bayesian networks have been proposed. The main question here is whether the results for synthetic data of such learning algorithms carry over to the real-world data of the multimorbidity domain. Therefore, it is necessary to know which of the learning algorithms are most suitable for this purpose. Compared to other domains, in general practice data there is large range in frequency of occurrence of particular events. Where in some domains structure learning algorithms that favour sparsity of connecting edges are more suitable, we do not know whether this holds for our domain. From a medical point of view each significant association that helps understanding the complex domain of multimorbidity would be valuable.

For example, cardiovascular disease prevalences were recently calculated in patients with inflammatory arthritis, osteoarthritis, and diabetes mellitus for a nationwide population in the Netherlands [141]. One of the results is that the prevalence of acute myocardial infarction ranges from 0.7% (controls) to 2.8% (diabetes mellitus). Although the absolute differences in prevalences are relatively low, the reported odds range up to 4. This suggests that treating one of these diseases is also of clinical importance in the light of cardiovascular risk management.

As described in some detail in Chapter 2, one class of the Bayesian network learning algorithms are the *search-and-score-based* methods. They attempt to identify a model that best fits the data by searching through the space of candidate models and selecting the one with the highest score. The search is guided by various heuristics, such as hill-climbing and tabu search. Typical scoring methods are the Bayesian in-

formation criterion (BIC), the Akaike information criterion (AIC), and the Bayesian Dirichlet equivalence (BDE) score. Another class of approaches, the *constraint-based* methods, estimate from the data whether certain conditional independences between the variables hold. Networks that are consistent with these independences are selected.

Bayesian network learning algorithms have several limitations. For example, search-and-score-based methods may end up in a local maximum. On the other hand, the constraint-based methods rely on the assumption that there exist no unobserved variables in the domain that explain (in)dependences between variables; this assumption almost never holds in reality. Nonetheless, it is well known that Bayesian networks learned from data typically perform quite well compared to simpler models, such as naive Bayes, when used for classification [121]. Furthermore, there have been significant efforts to ensure that the learned structure is correct [200]. Despite all the work on new learning algorithms, little is known, however, about how well they perform on a real-world dataset in the domain described here.

We used the following, special methodology for the purpose of the research. In the well-understood domain of cardiovascular medicine, a Bayesian network was drafted manually, based on expert knowledge from clinical guidelines. The choice for this clinical problem domain ensured that we were not faced with the typical problem in medicine that interactions between variables were only partially known. The structure of this network was used to generate synthetic datasets. The real-world data we used comes from general practices in the Netherlands, the same as used in [141]. The large size of this dataset makes it very likely that the statistics of the data conform to those in the guidelines. We compared different structure learning algorithms on both the synthetic and real-world data. We chose to evaluate a selection of algorithms – representing constraint based, search-and-score, and hybrid algorithms – that are scalable and often used.

5.2.2 *Related work*

Evaluating the performance of different model selection criteria has been done before, see e.g., [205] and [226]. However the results vary. Scoring methods such as BIC, AIC, and BDE have their own advantages and disadvantages in specific settings. Other related empirical work by de Jongh and Druzdzel [36], who investigated structural evaluation measures for Bayesian networks rather than scoring functions. They concluded that the structural Hamming distance, explained below, is especially useful when looking for causal structures.

In most studies the data used in the research has been generated: either the training data, the ground truth data, or both are synthetic. Based on such studies it is hard to conclude anything about the usefulness of particular methods for the analysis of real-world data. More recently Liu, Malone, and Yuan [115], carried out an empirical evaluation of scoring functions for Bayesian network selection using datasets from the UCI machine learning repository [1], which are notoriously unreliable. Their results showed that the BIC outperforms several other scoring measures, including AIC and BDE. However, as the biomedical datasets – our field of interest – in this repository

1 http://archive.ics.uci.edu/ml/

have only a limited number of patients (between 100 and 700), they generated larger datasets by sampling from these datasets.

A comparison of several structure learning methods for Bayesian networks, on a large medical dataset, was done by Acid et al [1]. They used several performance measures, e.g., the Kullback-Leibler distance, to evaluate how well the learned models fitted the data. Their research is closely related to our work. We also compared constraint based, search-and-score, and hybrid algorithms. The major difference is that in our research we compare such algorithms in a more qualitative manner by comparing the learned structures with the true underlying structure (assuming this is known).

Other work on evaluating Bayesian network scoring criteria was done by Jiang et al. [88, 89], where Bayesian networks were used to learn genetic interactions. In this medical domain, it is assumed that most of the *single* genetic variables (called SNPs) *are not* correlated with a disease, but that in contrast *combinations* of them *are* correlated. Although dealing with genetic data involves dealing with extreme low probabilities as well, the condition that single variables cannot explain a disease does not hold in the multimorbidity domain, e.g., single diseases, such as hypertension or diabetes, are also correlated with the onset of other chronic diseases. Another difference is that, in genetics, the direction of the arcs in a Bayesian network are obvious ($\mathtt{gene}\ \mathrm{G} \rightarrow \mathtt{disease}\ \mathrm{X}$). In multimorbidity, however, this is not always the case. Either $\mathtt{disease}\ \mathrm{X} \rightarrow \mathtt{disease}\ \mathrm{Y}$, or $\mathtt{disease}\ \mathrm{Y} \rightarrow \mathtt{disease}\ \mathrm{X}$ may hold, and they may also interact in a way making it impossible to decide on the direction.

5.2.3 *Comparing Bayesian network structures*

Information about structural similarity and differences between two Bayesian networks is usually determined by a metric called the *structural Hamming distance*. This metric is defined in terms of additional (arcs in the learned network not present in the true network), missing (arcs in the true network not present in the learned network), and reversed arcs (arcs in the learned network that has the opposite direction).

Multiple Bayesian network structures with arcs in different directions may belong to the same equivalence class. Intuitively, the distance between Bayesian networks in the same equivalence class should be zero. An equivalence class is represented by a *partially directed graph* (PDAG), also called essential graph, in which some edges are directed and some undirected. The undirected edges can be orientated arbitrary as long as no new v-structure in which multiple variables share a child is introduced. The SHD then counts the number of directed and undirected edge additions, deletions, and reversals to transform one PDAG into the other as the distance between two corresponding Bayesian networks. For details on the SHD consult [200]. Besides the SHD, the Hamming distance on the skeletons were also calculated. Skeletons are undirected and therefore only involves counting additions and deletions.

5.2.4 *Limitations of structure learning*

There are some limitations of Bayesian network structure learning. For example, the constraint-based methods rely on the assumption that there are no unobserved vari-

Sample		Grow	Tabu Search			MMHC		
Size	p_i	Shrink	BIC	AIC	BDE	BIC	AIC	BDE
1000	0.25	62	55	78	73	55	84	73
	0.10	8	3	27	11	3	28	10
	0.05	2	1	11	4	1	6	2
10,000	0.25	94	99	89	98	99	94	98
	0.10	58	32	77	53	32	84	53
	0.05	8	2	26	5	1	30	5

Table 5.1: Structure learning performance of the structure $u \rightarrow v \leftarrow w$ for different sample sizes and different probabilities distributions as in equation 33. Shown is the percentage of correctly identified structures for 1000 simulations.

ables in the domain that explain (in)dependences between variables. This assumption is almost never satisfied. For example, in the general practice data we used only measurements that were present in the data. Very often in clinical data tests with a negative result are missing.

Score-based methods are less sensitive to the assumption of no hidden variables. However, from previous studies one can conclude that search-and-score-based methods have the tendency to construct graphs that are sparser than the true underlying Bayesian network structure. Schulte et al. [181] showed that even for the simple three-vertex network $u \rightarrow v \leftarrow w$, the BDE scoring method does not find the true structure in over 50% of the cases for uniform randomly generated distributions and sample sizes below 1000.

We repeated this experiment for binary variables with a very simple binomial distribution as follows:

$$\begin{aligned} u &\leftarrow \text{binomial}(p = 0.30) \\ w &\leftarrow \text{binomial}(p = 0.20) \\ v &\leftarrow \text{binomial}(p = 0.10 + p_i \cdot u + p_i \cdot w) \end{aligned} \tag{33}$$

If we run this experiment a 1000 times for different values of p_i and two different sample sizes, we obtain the results in Table 5.1. One can see that for lower values of p_i and for a lower sample size the performance decreases rapidly. The search-and-score-based methods using the AIC performed the best. Altering the probabilities of u and w, or the type or error-size (α) of the conditional independence tests, lead to the same conclusions. This is a major concern since the risk attributions, for $p_i = 0.05$ in our case, still imply an odds ratio of 2 when both conditions are present. In medicine such a result is still considered of clinical value.

Although the gap between the GS algorithm – as second best – and search-and-score-based algorithms using the AIC as score is quite large, the GS algorithm performs equally well when learning the skeletons. The major problem of the GS algorithm is the need to orientate the direction of edges. The other scoring methods (BIC and BDE) lead to more sparsity of the learned structures. In the following section we will explore this further for synthetic datasets and the real-world medical dataset.

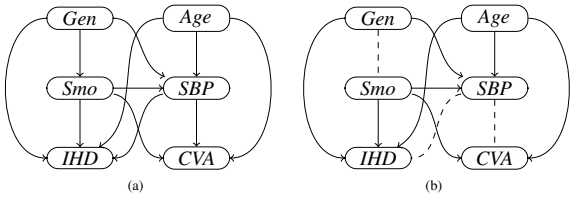


Figure 5.1: Bayesian network structure determined from a clinical guideline (a), which is used for the experiment with synthetic data, and its equivalence class (b). Dashed arcs represent the undirected arcs of the equivalence class.

5.2.5 *Structure learning from a large patient dataset*

The structure learning algorithms were applied to a synthetic dataset (SD) and a patient dataset (PD). The PD was extracted from the LINH data (see Chapter 3). We applied the algorithms to all possible clusters of three diseases taken from the diseases that were investigated in [141], i.e., inflammatory arthritis, osteoarthritis, diabetes mellitus, hypertension, hypercholesterolemia, acute myocardial infarction, transient ischemic attack, and cerebrovascular accident. The average number of significant edges identified by the GS algorithm was 2.0, however only 44% of these edges could be directed. Tabu search and MMHC using the AIC identified an average of 2.2 and 2.0 respectively (all directed). These numbers drop to 1.5 when using the BIC or BDE.

In a second experiment, the Dutch clinical guideline on cardiovascular risk management [223] was examined thoroughly to determine a network with cardiovascular disease related variables. Figure 5.1 shows the model that was used to generate the synthetic dataset, and which is assumed to be the true model of the dependences and independences between the variables age (Age), gender (Gen), smoking (Smo), systolic blood pressure (SBP), ischemic heart disease (IHD), and cerebrovascular accident (CVA). This simple model already shows a high density of associations between the disease variables. In fact, the skeleton is already close to a fully connected graph. However, it is assumed that IHD and CVA are more or less independent given age, gender, smoking, and systolic blood pressure. Other factors may be necessary as well to obtain conditional independence, e.g., lipid profiles, but were not used in our approach. The guidelines states that gender has a significant role in IHD. For CVA this is far less significant and therefore an arc between them is not present in the model.

In the synthetic dataset, gender was binomially distributed with a probability of 50%. The variable age was uniformly distributed between 40 and 100 years. Smoking was binomially distributed with a probability of 20%, being a female adds another 1%. The SBP was normally distributed with a mean of 120 and a standard deviation of 10. However, for each 10 years above 40, or being a male, or being a smoker, the mean was increased with 5, making the true mean of SBP approximately between 135 and 140. IHD and CVA were binomially distributed, both depending on age and SBP, where IHD depended more on age, and vice versa CVA more on SBP. For males 5% was added to the probability of IHD. For smokers 5% was added to both the probabilities

of IHD and CVA. These distributions are comparable with the distributions found in the PD when the variables are analysed with general linear regression methods. The following equations summarise the distributions describe above:

$$
\begin{aligned}
\text{gen} &\leftarrow \text{binomial}(\text{p} = 0.50) \\
\text{age} &\leftarrow \text{uniform}(\text{interval} = [40, 100]) \\
\text{smo} &\leftarrow \text{binomial}(\text{p} = 0.20 + 0.01(\text{gen} = \text{female})) \\
\text{SBP} &\leftarrow \text{normal}(\text{mean} = 120 + 0.5(\text{age} - 40) \\
&\qquad + 5(\text{gen} = \text{male}) + 5(\text{smo} = \text{yes}), \text{sd} = 10) \\
\text{IHD} &\leftarrow \text{binomial}(\text{p} = 0.05(\text{gen} = \text{male}) + 0.002 \cdot \text{age} \\
&\qquad + 0.05(\text{smo} = \text{yes}) + 0.001(\text{sbp} - 120)) \\
\text{CVA} &\leftarrow \text{binomial}(\text{p} = 0.001 \cdot \text{age} + 0.05(\text{smo} = \text{yes}) \\
&\qquad + 0.002(\text{sbp} - 120))
\end{aligned}
$$

The same variables were extracted from the PD for patients older than 40 years. This left us with approximately 150,000 patients, for which the SBP represents an average of all SBP measurements over a period of five years. Both the age and the SBP are discretised in intervals of ten.

For both datasets Bayesian network structures were learned using the following algorithms: the GS algorithm; TABU with the scoring functions BIC, AIC, and BDE; and the MMHC algorithm, again with the same scoring methods. The experiments were repeated for datasets containing 250, 1000, 10.000, and 100.000 sampled patients respectively. For each sample size we sampled a thousand times and calculated the average SHD of the PDAGs and Hamming distance of the skeletons. Eventually the structure learning algorithms were applied to the entire PD, see Table 5.2.

In the resulting models of the experiments, it appeared that in the PD the arcs $gender \rightarrow age$ and $age \rightarrow smoking$ were learned systematically in most of the models, even for low sample sizes. Figure 5.2 shows the evolution of the model for increased sample sizes. A close analysis of the PD for these variables (not shown here), shows that females are on average one year older than males, and that non-smokers are on average five years older than smokers. After adding these arcs to the true model, Table 5.3 shows the newly calculated Hamming distances of the sampled real-world data from the PD. In the learned model for the overall population, using MMHC with AIC, the only difference with this adjusted model is that instead of an arc SBP $\rightarrow$ CVA there is an arc IHD $\rightarrow$ CVA.

For the synthetic dataset, the tabu-search algorithm with the AIC as scoring function performed best for smaller datasets. With larger sample sizes, it was outperformed by the MMHC. The same conclusion held for the real-world datasets, except for determining the skeleton when the sample size was large. In that case, the GS algorithm outperforms the search-and-score-based methods. Still, even when the skeleton as discovered by GS was given, it had difficulty in orientating the arcs properly. Again, altering some of the parameters in the learning methods (α, restarts, perturbations), led to the same conclusions (not shown here).

Sample Size	Dataset	Hamming Distance	Grow Shrink	Tabu Search BIC	Tabu Search AIC	Tabu Search BDE	MMHC BIC	MMHC AIC	MMHC BDE
250	synthetic	skeleton	9.10	10.7	**7.95**	8.83	10.8	9.07	9.48
		structural	11.1	11.0	**10.9**	11.3	11.0	11.1	11.1
	real-world	skeleton	10.1	10.1	**9.01**	9.28	10.2	9.83	9.98
		structural	11.7	11.1	**11.0**	11.6	11.1	11.5	11.5
1000	synthetic	skeleton	7.83	8.95	**5.49**	6.67	9.31	7.60	8.09
		structural	10.2	11.0	**9.91**	10.9	11.0	11.0	11.0
	real-world	skeleton	9.33	9.50	**8.17**	9.19	9.64	9.07	9.45
		structural	11.8	11.8	**10.8**	11.4	11.6	11.1	11.3
10,000	synthetic	skeleton	6.26	5.25	**3.01**	4.67	5.63	4.72	5.27
		structural	8.19	10.6	**5.13**	8.04	10.6	7.19	8.48
	real-world	skeleton	**7.04**	7.76	7.46	7.22	8.31	7.25	7.58
		structural	12.0	10.5	10.2	10.4	10.2	**9.13**	9.28
100,000	synthetic	skeleton	2.47	2.34	1.56	2.08	2.39	**1.53**	2.25
		structural	4.62	5.32	2.11	4.71	5.82	**1.55**	5.36
	real-world	skeleton	**4.12**	7.44	5.51	8.00	7.44	4.69	8.00
		structural	10.8	9.89	8.69	11.8	9.44	**6.03**	10.0
154115	real-world	skeleton	**4**	8	5	8	8	**4**	8
	(complete)	structural	11	10	6	13	10	**5**	10

Table 5.2: Average Hamming distances of the learned Bayesian network structures (row-lowest values are in bold, MMHC=max-min-hill-climbing).

Whereas Liu et al. [115] report good results using the BIC for relatively small datasets, we could not confirm their results using our real-world data. The AIC performs best for small datasets with low marginal probabilities. This was expected as the datasets from the UCI repository have distributions that differ in their characteristics from the datasets we studied, in particular for our datasets there may be very low probabilities for particular events. The sensitivity of structure learning methods to extreme probabilities deserves a more thorough analysis, as this is something that occurs in many medical domains.

A further observation is that the difference between the distance of the skeleton and the distance of the PDAG to the true model is fairly constant for the synthetic dataset when increasing the sample sizes. This applies more or less to all algorithms used here. However, for the real-world data this is not true, i.e., most of the algorithms failed to discover the correct direction of arcs when increasing the sample sizes. The only exception was the MMHC algorithm with an AIC score, where the differences remained fairly constant.

Some of the differences to the literature we found can be explained by the fact that our models are more dense compared to randomly generated networks. In real-world patient data it is hard to make comorbid diseases independent since there are several confounders such as laboratory measurements and life style factors. These are typically

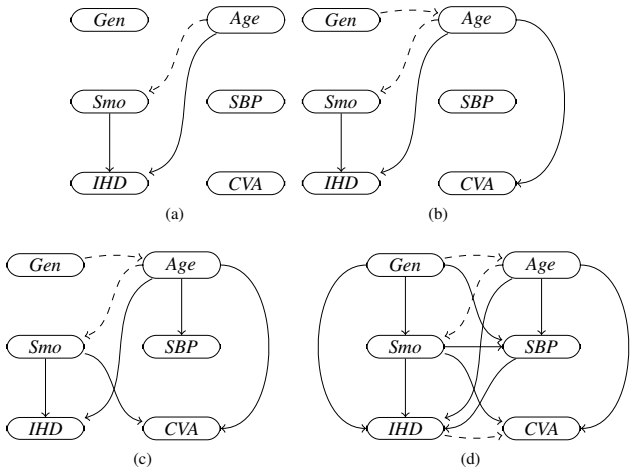


Figure 5.2: Evolution of the learned model for the real-world dataset using the MMHC algorithm with the AIC score. The models (a), (b), (c) and (d) are learned with a sample size of 250, 1000, 10000 and 154115 (the complete dataset), respectively. Dashed arcs represent arcs that were not present in the original model in Figure 5.1.

not all known in epidemiological data, which leads to even more dense networks. Another issue is that the true direction of arcs between such confounders is often not clear, which made it difficult to use them in this evaluation.

5.3 MODELS OF CO-OCCURRENCE AND THEIR CHARACTERISATION

In the previous section, we determined that, in case of a large patient dataset with low disease prevalences, using a *search-and-score-based* method with a score that depends on the AIC, performs well for learning the structure of a Bayesian network. Therefore, we use this method to learn the structure of Bayesian network models for explaining the observed associations between oncologic diseases. Before moving on to the characterisation of such networks, we first briefly recall the measures of association frequently being used in multimorbidity research.

5.3.1 *Statistical measures of association in multimorbidity*

Commonly used measures in medicine to describe associations are the relative risk (RR) for disease pairs and the ϕ-correlation coefficient. These measures have been used to investigate cancer metastasis patterns in a network-based manner, where edges in a constructed network were added because of high enough strength of RR or ϕ-correlation [28].

For readability, we recall the definition of the RR for disease pairs, which coincidences with the multimorbidity coefficient. Let N_i be the number of patients with

Sample	Hamming	Grow	Tabu Search			Max-min Hill-Climbing		
Size	Distance	Shrink	BIC	AIC	BDE	BIC	AIC	BDE
250	skeleton	10.8	12.0	**9.09**	10.5	12.1	10.7	11.0
	structural	12.9	13.1	**12.1**	13.0	13.0	12.8	13.0
1000	skeleton	8.99	10.2	**7.13**	9.21	10.4	8.57	9.71
	structural	12.1	13.1	**10.7**	12.2	13.0	11.4	12.4
10,000	skeleton	**5.08**	5.84	5.46	5.22	6.39	5.25	5.58
	structural	11.5	10.0	8.84	9.46	9.41	**7.35**	7.87
100,000	skeleton	**2.12**	5.44	3.51	6.00	5.44	2.69	6.00
	structural	9.76	8.08	7.73	10.4	7.44	**3.16**	8.00
154,115	skeleton	**2**	6	3	6	6	**2**	6
(complete)	structural	10	8	3	10	8	**2**	8

Table 5.3: Adjusted average Hamming distances of the learned Bayesian network structures (row-lowest values are in bold).

disease D_i, N_j the number of patients with disease D_j, N_{ij} the number of patients with both diseases D_i and D_j, and N the total number of patients. The *relative risk* of observing a pair of diseases D_i and D_j affecting the same patient is then given by

$$RR_{ij} = \frac{N_{ij}N}{N_i N_j} = \frac{P(d_i, d_j)}{P(d_i)P(d_j)} \tag{34}$$

The statistical significance of the RR depends on the sample size, the size of the prevalences involved, and the noise in the sample. Further characteristics of the RR and its use in medicine are outlined in [190]. If we evaluate the RR for a specific subpopulation by *conditioning* on a set of risk factors Q, we obtain:

$$RR_{ij}^{q} = \frac{N_{ij}^{q}N^{q}}{N_i^{q}N_j^{q}} = \frac{P(d_i, d_j \mid q)}{P(d_i \mid q)P(d_j \mid q)} \tag{35}$$

with N_{ij}^{q} the absolute prevalence of both D_i and D_j within the subpopulation of patients for which $Q = q$ holds. N^{q}, N_i^{q}, and N_j^{q} are defined likewise.

There are only a few situations in which multiple diseases occur together. The first possibility is that their co-occurrence is at random, i.e., exactly from what can be expected by chance. Adopting the standard probabilistic terminology, this notion is called *independent multimorbidity*, see also Chapter 4, and it coincidences with an RR $= 1$. The opposite of independent multimorbidity is called *associative multimorbidity*, which means that there is some relationship between diseases which causes the co-occurrence of the diseases to be different from expectation by chance. Typically, one is interested in a *positive* associative multimorbidity, i.e., where $RR > 1$.

While independence measures provide some insight into which diseases might co-occur more frequently, they do not give much insight into the aetiology as the are a number of ways these diseases can be related [102]. To model and analyse such relationships between diseases, we will use Bayesian networks, as these can be used to

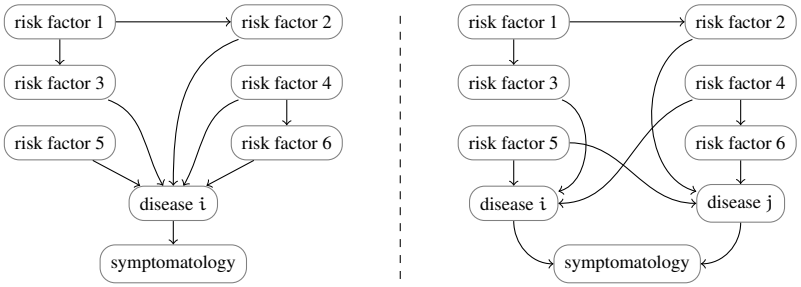


Figure 5.3: Bayesian network of a single disease (left) and multiple diseases (right).

model more complex structural relationships between disease variables in comparison to traditional regression models.

5.3.2 *Structural measures of multimorbidity*

To illustrate how Bayesian network modelling can contribute to understanding the relationship between diseases, consider Figure 5.3, which shows a Bayesian network of a single and a multiple disease model. In both models there is a set of risk factors present, which can be any subset of environmental, patient, genetic, and other disease related variables. Between these risk factors (in)dependency can occur. For example, in both disease models, the risk factors 2 and 3 are associated with each other through a third risk factor 1, and the risk factors 4 and 6 are directly associated. The risk factor 5 is independent from the other risk factors, however in the multiple disease model it is also a common parent of both diseases.

There are a number of characteristics considered to be relevant in multimorbidity research, see also Chapter 4, namely whether the diseases are: (1) causally related (*direct causation model*), (2) related because of common risk factors (*heterogenic risk factor model*), or (3) the diseases are related because of risk factors that are correlated (*associated risk factor model*). Learning causal models is beyond the scope of this chapter, so we will focus on characteristics of heterogenic and associated risk factors of diseases. In the example above, risk factor 5 is a common risk factor, while, for example, risk factor 3 and risk factor 6 are heterogenic risk factors. The quantitative measures that we will use in this section, are the number of common risk factors and the number of associated risk factor combinations that lead to associations between diseases. As these measures show the number of relationships between the most important risk factors, they given an indication of the *complexity* of the reason for co-occurrences. If both numbers are 0 and their is no direct path between the diseases, then we have the simplest type of multimorbidity, i.e., independent multimorbidity.

5.3.3 *Critical factors*

While there can be many associations between diseases, often these diseases can be explained by only a few risk factors. For example, in Figure 5.3, risk factors 1, 4, and 5

completely explain the association between the diseases. If these risks can be prevented or reduced through an intervention, both the chance of the occurrence of the individual diseases *and* the multimorbidity burden is reduced, so these are the most *critical factors* in the model. More formally, this means we are interested in sets of factors C such that, given two diseases D_i and D_j and a graph G:

1. all trails $D_i \leftarrow X_1 \leftarrow \cdots \leftarrow X_k \rightarrow \cdots \rightarrow X_n \rightarrow D_j$ in the graph G are d-separated by C, i.e., $C \cap \{X_1, \ldots, X_n\} \neq \varnothing$;
2. there is no $C' \subset C$ for which the previous condition holds.

It is easy to see that there exists such a C such that every $X \in C$ is an ancestor of both D_i and D_j. In our non-causal models, the ultimate causes are not necessarily ancestors, hence, we take any of the minimal separating sets, which can be can be found in polynomial time [199].

Using existing techniques, it is possible to find a conditioning set Q such that $\forall q \in Q : RR_{ij}^{q} = 1$. If $\forall q \in Q : RR_{ij}^{q} = 1$, using Equation 35 this implies that:

$$\frac{P(D_i, D_j \mid Q)}{P(D_i \mid Q)P(D_j \mid Q)} = 1$$

so therefore we have:

$$\frac{P(D_i, D_j \mid Q)}{P(D_i \mid Q)P(D_j \mid Q)} = 1 \Leftrightarrow D_i \perp\!\!\!\perp D_j \mid Q \Leftrightarrow Q \text{ d-separates } D_i \text{ and } D_j \quad (36)$$

For example, the RR of a colorectal cancer and a respiratory cancer being a comorbid combination is 5. If we condition on the presence of liver cancer, the RR drops to 1.3 and it remains 5 when liver cancer is not present. In this approach, no distinction can be made between direct causation and common risk factors, i.e., whether liver cancer is a common risk factor or whether it is a metastasis of a colorectal cancer that will further metastasise to the lungs. In Bayesian network structure learning, it is common practice to include background knowledge during the learning of networks, e.g., knowledge that metastatic spread of a cancer is common. Furthermore, in structure learning approaches it is possible to use model selection that gives an indication of the best possible model, which is significantly more difficult using relative risk or Pearson's correlation.

5.3.4 *Experiments*

Recently, the LINH data was used to compare the occurrence of pre-existing and subsequent comorbidity among older cancer patients with older non-cancer patients in terms of odds and hazard ratios [38]. Here we explore the *co-occurrence* of cancers and their related comorbidity. In the remainder of this chapter we will denote malignant diseases, e.g., skin cancer or breast cancer, by M_i. Remaining chronic, but benign, conditions, such as chronic liver disease or benign prostatic hypertrophy, are denoted by C_k.

Table 5.4 shows prevalences of the most significant comorbid combinations of malignant tumours, corrected for age and gender. A combination of two malignant tumours M_i and M_j is selected when $M_i \not\perp\!\!\!\perp M_j \mid \{Age, Gender\}$, with a significance level

< 0.05. For comparison, the RR and the ϕ-correlation coefficient are calculated as well.

For each malignant tumour M_i present in Table 5.4, the set R_i consists of all associated conditions C_k for which $M_i \not\perp\!\!\!\perp C_k \mid \{Age, Gender\}$, with a significance level < 0.05, shown in Table 5.5. A distinction is made between disorders that are associated with the onset of a cancer, e.g. smoking, alcohol abuse, or a chronic liver disease (Table 5.5(a)), and disorders that are probably a consequence of a cancer, e.g., anaemia, depression, or cardiovascular disease (Table 5.5(b)).

Finally, we applied structure learning, using the R statistical software package bnlearn [185], for each $\{M_i, M_j, Age, Gender\} \cup R_i \cup R_j$ of each combination M_i and M_j present in Table 5.4. Within each structure we determined the minimal d-separation between M_i and M_j, and whether a direct association still remained between the two malignant tumours M_i and M_j. A pair of two associated risks R_k and R_l can be dependent or independent when corrected for age and gender, i.e., $R_k \not\perp\!\!\!\perp R_l \mid \{Age, Gender\}$ or $R_k \perp\!\!\!\perp R_l \mid \{Age, Gender\}$, respectively. Dependent risk factors can be grouped into common parents and associated risks. The results are showed in Table 5.6.

In our results, we observe that in the majority of cases there is no direct arc between two malignant tumours M_i and M_j if the $RR_{ij} < 10$, i.e., the association can only be explained by a set of critical factors. Age and gender frequently act as a common parent. Therefore, they are often part of the critical factors. Only on three occasions a disease variable, other than age or gender, acts as a direct common parent: chronic liver disease as common parent of pancreatic cancer and liver cancer; smoking as a common parent of respiratory cancer and bladder cancer; and benign prostate hypertrophy as a common parent of prostatic cancer and bladder cancer.

Figure 5.4a shows the local network structure that connects colorectal cancer and respiratory cancer. These two malignant tumours do not share a direct common parent, and there is no edge between these two variables in the network. However, the aetiologic association can be explained by the critical factors and the elements of directed paths, in this case: *age*, *smoking*, *alcohol abuse*, and *liver cancer*. Conditioning on these variables should lower the RR significantly. Indeed, if we condition on the facts that a patient smokes, has liver cancer, and is of age between 65 and 80, the RR observing both cancers drops from 5 to 1.8.

Figure 5.4b shows the local network structure that connects liver cancer and pancreatic cancer, which also appears at the top of the list in Table 5.4b. There is a direct association between liver cancer and pancreatic cancer. In this case, this is most likely a direct causal pathway, i.e., it is due to metastasis of pancreatic cancer to the liver. The remainder of the aetiology is totally explained by the direct common parents *age* and *chronic liver disease*. Colorectal cancer is associated with liver cancer, but as risk factor it is independent of pancreatic cancer.

5.4 CONCLUSIONS

The advantage of studying multimorbidity by means of Bayesian networks is that they allow modelling relationships between multiple disease variables, whereas it is also possible to learn these networks, to a large extent, from data. Structure learning has been applied before to medicine, for example, to predict disease related mortality [31],

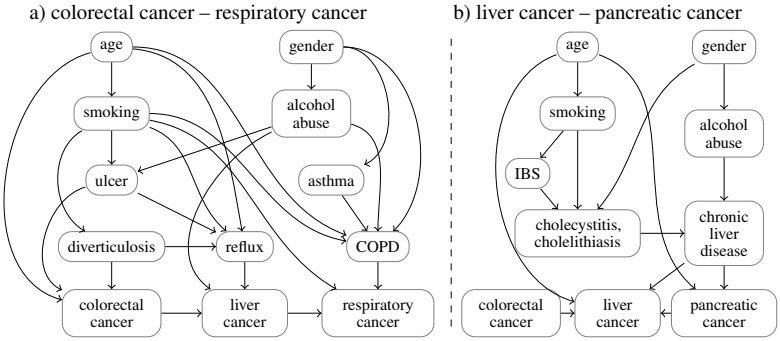


Figure 5.4: Bayesian networks for comorbid combinations of cancer. Abbreviations: COPD = chronic obstructive pulmonary disease; IBS = irritable bowel syndrome.

and also in the context of multimorbidity, e.g., in genome-wide association studies to determine genetic links between chronic diseases [3]. In this chapter, we use these methods to explain the co-occurrence of multiple diseases in terms of their risk factors. As the types of relationships between risk factors varies, we proposed looking at the critical factors that lead to co-occurrences.

We discussed the characteristics of disease networks, in particular that one has to deal with low prevalences and a dense structure. We showed that this has certain implications for the methods that are suitable for learning a Bayesian network structure. In order to do this, we assessed several well-known BN structure-learning algorithms. The aim was to obtain insight into whether the experimental results from the literature, which mostly are based on synthetic data or low-quality datasets from UCI, carry over to a real-world dataset when it comes to analysing multimorbidity. For that purpose, we used a very large real-world dataset, acquired from general practices, and focused on the well-understood domain of cardiovascular medicine. It appears from the results that structure learning is fairly sensitive to extreme probabilities.

In a further evaluation, we applied structure learning methods to a dense model of comorbid cardiovascular diseases with a relatively low average prevalence. In contrast to the synthetic data with similar characteristics, we found that a constraint-based algorithm outperformed the other methods when trying to find the correct skeleton of the network if the sample size is large. This suggests that research results obtained from synthetic data—quite typically something used to evaluate new structure-learning algorithms—do not necessarily carry over to real-world data.

The full Bayesian network can best be discovered using a search-and-score-based algorithm with the AIC as score method. While use of the AIC score may give rise to inclusion of too many arcs [115], from a medical point of view, we can argue that it is better to find all relevant associations, as this may significantly influence the patient's risk profile when making predictions based on the network. Although our database contains data from a nationwide population, generalising the results would be too premature. In future work we aim to further validate these results.

Another question is whether the learning algorithms can be improved. For example, the MMHC method uses a restricted search space based on learned skeletons. Since

the skeletons learned by the GS algorithm outperform the rest, one might replace the restricted search space by the skeletons obtained by the GS algorithm. In some preliminary experiments, however, we were unable to detect significant improvements using this idea. Nevertheless, we believe that there is a need for further improvement of structure-learning algorithms to deal with large real-world medical datasets.

Using a search and score method, we learned Bayesian network structures for pairs of oncologic diseases and their related comorbidity. The resulting networks are in line with knowledge known from the oncology literature. For example, long sustained exposure to chronic conditions, e.g., unhealthy lifestyles, is associated with high prevalence of breast, colorectal, respiratory and prostate cancer [219]. In the networks learned here, age and gender are often a direct parent of both malignant tumours. Other critical risk factors, e.g., smoking and alcohol abuse, are often associated by directed paths of pathophysiology to pairs of malignant tumours.

In some cases a direct edge remains in the network between the two malignant tumours. Most probable, this reflects metastasis, however direct associations found between two malignant tumours may also have another explanation, e.g., a genetic predisposition or another unknown confounder. Sometimes there a directed path of malignant tumours. This might reflect metastasis to secondary locations, e.g., *colorectal cancer* $\rightarrow$ *liver cancer* $\rightarrow$ *respiratory cancer*. The same holds for associations between risk factors, such as oesophageal reflux with diverticulosis, which may be explained by life style factors or genetics that are not present in the data.

The methodology used here shows that, even though overall relative risks between pairs of malignant tumours can be high, a direct association, such as metastasis, is not always the obvious reason for that. Using structure learning and concepts such as d-separation in Bayesian networks we identified other critical risk factors, e.g., age, gender, smoking, and alcohol abuse, in the pathogenesis of co-occurring malignant tumours. This shows that the method provides useful results for identifying critical factors of associated comorbid diseases where the role of such risk factors is less obvious.

cancer	prevalence
skin	21.33 ‰
breast	12.11 ‰
colorectal	7.43 ‰
respiratory	6.23 ‰
prostate	5.94 ‰
bladder	2.83 ‰
liver	2.64 ‰
muscles	2.40 ‰
leukaemia	2.13 ‰
female genitals	2.02 ‰
lymphoma	1.70 ‰
cervix	1.10 ‰
stomach	1.08 ‰
kidney	1.08 ‰
pancreas	1.07 ‰
neurologic	0.91 ‰
metabolic	0.59 ‰
ureter	0.25 ‰

(a) Prevalence

cancer site 1	cancer site 2	p-value CD-test	RR	ϕ
pancreas	liver	7.31E-16	34	0.22
respiratory	neurologic	1.80E-15	22	0.17
respiratory	liver	2.63E-14	10	0.13
stomach	liver	7.52E-13	33	0.13
pancreas	metabolic	6.62E-10	80	0.22
leukaemia	lymphoma	6.86E-10	21	0.41
bladder	kidney	1.03E-07	27	0.28
colorectal	liver	1.09E-07	8	0.13
bladder	ureter	1.84E-07	72	0.41
skin	lymphoma	5.66E-06	5	0.12
colorectal	respiratory	4.19E-04	5	0.09
breast	respiratory	8.75E-04	3	0.14
muscles	respiratory	1.32E-03	3	0.55
colorectal	kidney	1.67E-03	7	0.37
cervix	female genitals	3.04E-03	17	0.55
bladder	respiratory	3.20E-03	7	0.16
bladder	prostate	3.21E-03	9	0.32

(b) Co-occurrence

Table 5.4: Prevalence and co-occurrence of malignant tumours. All RRs and ϕ correlation coefficients are significant ($p < 0.05$). CD=conditional dependence.

System	Conditions
Patient	age, gender, smoking, alcohol, obesity
Digestive	oesophageal reflux, ulcer, chole-cystitis/lithiasis, viral hepatitis, chronic liver disease, irritable bowel syndrome, diverticulosis
Metabolic	diabetes mellitus, lipid disorder, hypothyroidism, hyperthyroidism, gout
Musculoskeletal	osteoporosis, rheumatoid arthritis, osteoarthritis
Neurologic	benign neurocancer, congenital anomaly
Respiratory	asthma, chronic obstructive pulmonary disease
Skin	acne, psoriasis, eczema, benign skin cancer
Urogenital	benign prostatic hypertrophy, endometriosis

(a) Associated Risks

System	Conditions
Cardiovascular	hypotension, hypertension, ischemic heart disease, heart failure, arrhythmia, myocardinfarct, stroke, embolism, varicosis, flebitis
Musculoskeletal	hernia, spondylosis, tendinitis
Neurologic	headache, epilepsy, neuropathy
Psychiatric	anxiety, personality disorder, depression, organic psychosis, somatisation, neurasthenia
Blood/Renal	anaemia, purpura, renal insufficiency
Eye/Ear	macula degeneration, cataract, deafness, vertigo

(b) Associated Symptomatology

Table 5.5: Chronic conditions associated with at least one of the cancers listed in Table 5.4, grouped by risk factors and symptomatology.

cancer site 1	cancer site 2	DP	RF	CP	AR	IR	critical factors (min. d-separation)
pancreas	liver	yes	12	2	0	1	age, chronic liver disease
respiratory	neurologic	yes	12	0	0	4	gender
respiratory	liver	yes	15	1	3	1	age, oesophageal reflux, chr. liver disease
stomach	liver	no	10	1	4	2	age, alcohol abuse, chronic liver disease
pancreas	metabolic	yes	8	1	0	0	age
leukaemia	lymphoma	yes	3	2	0	0	age, gender
bladder	kidney	yes	10	1	1	0	age
colorectal	liver	yes	13	1	2	0	age, diverticulosis
bladder	ureter	yes	6	1	0	0	age
skin	lymphoma	no	15	1	0	3	age
colorectal	respiratory	yes	15	1	3	0	age, smoking, alcohol abuse
breast	respiratory	no	15	2	4	0	age, gender, smoking, osteoporosis
muscles	respiratory	no	13	0	3	3	asthma, arthrosis
colorectal	kidney	no	11	1	1	1	age, diverticulosis
cervix	genitals	yes	3	1	0	0	gender
bladder	respiratory	no	14	3	2	0	age, gender, smoking, benign skin tumour
bladder	prostate	no	14	3	1	0	age, gender, benign prostate hypertrophy

Table 5.6: Aetiological measures of comorbid malignant tumours. Abbreviations: DP = directed path, RF = the number of risk factors in the local network, CP = the number of common parents, AR = the number of risk combinations that are associated, IR = the number of risk combinations that are independent.

6

MULTILEVEL BAYESIAN NETWORKS FOR THE ANALYSIS OF HIERARCHICAL HEALTHCARE DATA

ABSTRACT

Large healthcare datasets normally have a hierarchical structure, that is captured in terms of 'levels', as the data may have been obtained from different practices, hospitals, or regions. Multilevel regression is the technique commonly used to deal with such multilevel data. However, for the statistical analysis of interactions between entities from a domain, multilevel regression yields little to no insight. While Bayesian networks have proved to be useful for analysis of interactions, they do not have the capability to deal with hierarchical data. In this chapter, we describe a new formalism, which we call multilevel Bayesian networks; its effectiveness for the analysis of hierarchically structured healthcare data is studied from the multimorbidity perspective. The results are compared with those obtained by multilevel regression. Using multilevel Bayesian networks we were able to obtain models with higher predictive power and with a significant net reclassification improvement. The Bayesian-network models offered considerable more insight into the interactions between the diseases through their structure. Moreover, a multilevel Bayesian network model can be used for the prediction of the occurrence of multiple diseases, even when some of the predictors are unknown. Thus, multilevel Bayesian networks offer an attractive alternative to multilevel regression equations when analysing hierarchical healthcare data.

6.1 INTRODUCTION

Healthcare research is often done using clinical data that have a hierarchical structure – they have *levels* as is said. This may be due to the fact that the data have been obtained from different practices, hospitals, or regions. Since patients within the same practice are often more alike than two randomly chosen patients, they will likely have some correlation on variables related to the practice. Statistical analyses that ignore these correlations will lead to results that are statistically invalid [167]. Commonly used statistical techniques such as logistic regression do not allow incorporating the characteristics of the different levels in the hierarchy. Therefore, multilevel regression methods are often used to analyse such data. The books [9] and [83] offer an overview of such methods.

In the artificial intelligence literature, the use of probabilistic graphical models, such as Bayesian networks [149], have had a significant impact on the modelling and analysis of patient data [120]. The edges in the graphical model represent probabilistic relationships between specific patient variables for a disease of interest. Bayesian networks allow for the integration of medical domain knowledge, and clinical expertise can be modelled explicitly. Moreover, clinical knowledge derived from clinical healthcare data can be used to further refine and validate the model.

In this chapter, we combine *multilevel* modelling and learning with Bayesian network modelling. This can be useful in complex domains such as the epidemiology of *multimorbidity* (see Chapter 3 for details). Multimorbidity is often analysed using multilevel regression, as it requires a large amount of data coming from different sources in order to study the interaction between diseases. Moreover, it is a typical problem where Bayesian networks can be useful, as expert knowledge is needed, and representing multiple diseases requires scaling up to models containing a large number of variables.

Since Bayesian networks have already been successfully applied to model single diseases [120, 8, 51, 97, 118, 119, 216], and also for multiple diseases [72, 73, 145, 146, 148], the research question is whether and how it is possible to adopt the multilevel approach for Bayesian networks. In that way we would be able to explore complex healthcare data that is hierarchically structured using Bayesian networks with the advantage that, in contrast to multilevel logistic regression, models are obtained that offer a clear representation of the interactions between multiple diseases.

The main contribution of this chapter is that it introduces a new representation of multilevel disease models using Bayesian networks, which we call *multilevel Bayesian networks*. It has the advantage that it is at least as powerful as multilevel logistic regression, yet supports, in contrast to multilevel logistic regression, gaining new insights into the interactions between multiple diseases.

Using patient data from family practices in the Netherlands, we applied this framework to obtain a prediction model for multiple chronic diseases, namely diabetes mellitus and heart failure. The effectiveness of multilevel Bayesian networks has been studied by comparing the resulting model to the traditional models based on multilevel regression analysis.

6.2 RELATED RESEARCH

Multimorbidity is the healthcare problem where we focus on in this chapter, although multilevel Bayesian networks may have other applications as well. The problem of multimorbidity has been extensive reviewed in Chapter 3 and the reader is referred to this chapter for further detail. Here we only briefly summarise some relevant facts.

In most of the research in multimorbidity, the prevalence and significance of specific factors for predicting the presence or absence of specific diseases is typically determined by means of (multilevel) regression methods, where the variance of the observations is minimised with respect to a linear or logistic model. Where multimorbidity should be studied by exploring the interactions between diseases with associated signs and symptoms in their full generality, in practice current research explores this only in a very restrictive fashion. For example, there is some research where the prevalence

of multimorbidity has been studied in family practices [207, 53] by means of clustering methods [124]. Multimorbidity indices are another way to measure specific types of multimorbidity within a population [39]. These methods illustrate the size, impact and complexity of multimorbidity, but all of them give little insight into interactions between diseases.

Multilevel regression has many applications in the social sciences and in medicine; however, it was not especially designed to model multimorbidity [15, 40, 75]. In [140] complex hierarchical patient data were used to analyse the predictive value of cardiovascular diseases for hypertension and diabetes mellitus. Since both diseases are analysed separately, the results only give a preliminary view on correlations between cardiovascular diseases.

Various Bayesian network models for multiple disease have been developed since the beginning of the 1990s. Examples are Pathfinder [72, 73], Hepar II [146] and MUNIN [198]. They deal with multiple diseases, although all of them belonging to the same disease class. One of few existing exceptions is QMR-DT [189, 132], as it covers a broad subset of internal medicine. However, it was never meant for actual use. All these Bayesian network models have been constructed based on expert opinion and engineering background knowledge. They did only incorporate *known* disease interactions; they were not meant for uncovering *new* disease interactions. This explains why dealing with multilevel data was not seen as a problem. In this chapter we make an important step forwards in this respect, as Bayesian network models are learned in order to gain insight into the interactions between diseases. Without the capability to deal with hierarchical data, using multilevel methods, models are obtained that may yield predictions that are statistically unsound.

Bayesian networks have also been used in algorithms for learning patient-specific models from clinical data to compare mixed treatments and to predict disease progression [162, 217]. Somewhat confusingly, the adjective 'hierarchical' is also used in connection to Bayesian networks. For example, nested, hierarchical Bayesian networks allow one to define genetic models that can be reused [194]. Hierarchical Bayesian networks have also been proposed as an aggregating abstraction [70] that clusters variables closely related to each other. This all closely relates to object-oriented Bayesian networks [94], but there is no relationship to multilevel analysis where the hierarchy stands for nested data from different groups.

Eventually, one would like to have methods for the analysis of healthcare data that can handle multimorbidity, and have the ability to be personalised as well, i.e., allow entering observations of a patient into the probabilistic model and obtaining updated parameters that specifically account for that patient. Such personalised models help to generate specific advice that relates to the patient's health status. The probabilities of the underlying model could be extracted from existing clinical research or from available patient data, using a valid method that takes interactions between diseases into account.

To illustrate the types of relationships that can occur, the left-hand side of Figure 6.1 shows the typical relationships between variables for a single disease, whereas at the right-hand side the integration of multiple diseases into one graphical model is depicted. Representing multiple diseases in one model avoids redundancy of separate representations and has the advantage that it shows where diseases interact. Mutual dependences

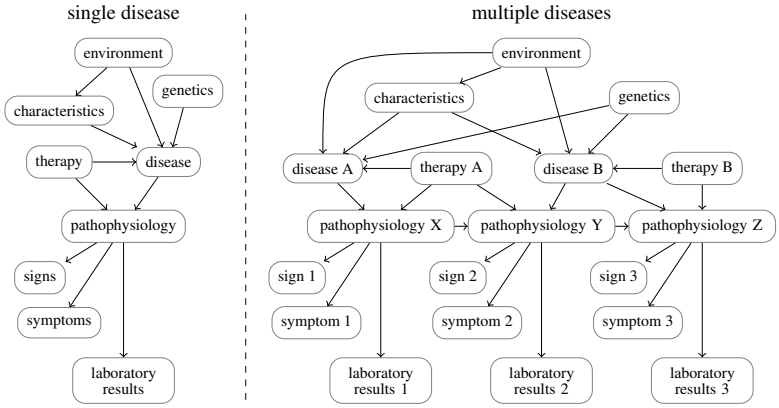


Figure 6.1: Abstract model of a single disease (left) and multiple diseases (right).

may concern diseases, therapies, pathophysiology, symptoms, signs, and lab results, and modelling interactions explicitly allows us to make better decisions for patients having multiple diseases. In fact, the architecture of networks such as MUNIN [198] is similar, as it also models diseases in terms of their pathophysiology and patient findings.

6.3 MULTILEVEL REGRESSION

To analyse multimorbidity problems one has to deal with large datasets in which variance is introduced by the fact that the data have been collected from different sources, such as family practices and populations, either social, economic, or demographic. If we would ignore this, identifying interactions between disease variables, such as pathophysiology and laboratory results, could be difficult and even erroneous.

While Bayesian networks model a joint probability distribution, regression methods estimate conditional distributions. Linear regression tries to estimate a linear dependency between the observations of a random continuous variable (assuming it is normally distributed), denoted by O, and a set of (non-random) *explanatory variables*, denoted by e. This is done by using an optimisation algorithm, such as the least square method, that minimises the deviation of the observations with respect to the model parameters.

If, additionally, the data is hierarchically structured, then at each level, the data can be split into *groups*. Characteristics of each group are modelled by additional (non-random) *level variables*, denoted by l. For example, if the different practices are modelled by a grouping variable, a variable such as urbanity that will be shared among practices is modelled by such a level variable. Multilevel analysis tries to explain the variance caused by level variables that have an influence on the explanatory variables e. For example, if we use linear regression, the intercept and slope, that determine the linear dependency between two variables, may alter for different groups.

More precisely, in multilevel regression we wish to explain an observation o with respect to explanations e and l, assuming that the observations o are possible outcomes of a random variable O. Let us first assume that there are only two levels to cope with the grouped data. The explanations e represent the first level, i.e., they can be different for each individual. The second level then represents the groups, which are characterised by the explanations l. The explanations l can thus only differ per group, and together with the explanations e they describe each individual.

Let there be r groups with n first-level explanations and m second-level explanations. Then, for each q^{th} group at the second level we define a linear regression model for O, and allow dependency of the regression coefficients on the variables l_j and certain deviation from the overall mean. With $e = (1, e_1, \ldots, e_i, \ldots, e_n)^T$, $l = (1, l_1, \ldots, l_j, \ldots, l_m)^T$, i.e., $n + m$ explanations, $\delta_q = (\delta_{0q}, \ldots, \delta_{nq})^T$ (the second level noise), for $q = 0, \ldots, r$, and β a matrix consisting of components β_{ij} (the effect of l_j on the explanation e_i), the model then becomes: $E[O_q \mid e, l] = (\delta_q + \beta l)^T e$, which, if the noise is normally distributed, can be interpreted as a conditional probability distribution:

$$P(O_q \mid e, l) \sim \mathcal{N}(\mu_q, \sigma) \tag{37a}$$

$$\mu_q = (\delta_q + \beta l)^T e \tag{37b}$$

for $q = 0, \ldots, r$, where the expectation of the outcome variable $E[O_q \mid e, l] = \mu_q$.

In this model, the outcome for each group is dependent of explanatory variables e weighed by the coefficients β, the level variables, and random variables δ_{iq}, where for each i, the δ_{iq} are normally distributed with expectation zero, and correlated with a $\delta_{iq'}$. These correlations ensure that observations for one group have an impact on other groups through this hierarchical structure.

Generally, multilevel models assume homogeneity of variance for all observations on the first level, i.e., σ is constant, and does not depend on e, l, and q. Likewise, it is also assumed that the variance on the second level is homogeneous, i.e., the variance of δ_{iq} is equal to σ_i^2, and the covariance of δ_{iq} and $\delta_{i'q}$ is equal to $\sigma_{ii'}^2$, and thus not group specific. But there is no reason why this should be true in all applications. An alternative is to allow heteroscedasticity, i.e., heterogeneity of variances among groups on at least one of the levels. Heteroscedasticity, however, requires additional modelling when estimating the different variances [20, 64, 96], and is not described in detail in this chapter.

Adding the observations l_j simply to the regression model as additional explanatory variables, i.e., $e = (1, e_1, \ldots, e_n, l_1, \ldots, l_m)^T$, with corresponding regression parameters, i.e., $\beta = (\beta_0, \beta_1, \ldots, \beta_n, \beta_{n+1}, \ldots, \beta_{n+m})^T$, we obtain a one-level regression model with $n + m + 1$ degrees of freedom, which corresponds to standard linear regression. The number of degrees of freedom in the multilevel model is $q(n + 1)(m + 2)$. Figure 6.2 compares standard regression and multilevel regression on a synthetic dataset with observations divided into two groups.

The concept can be extended to more levels, e.g., three levels. If the q subgroups can be grouped further into s meta-groups, we can define a three-level model, with $l_1 = (1, l_{21}, \ldots, l_j, \ldots, l_{2m_1})^T$, and $l_2 = (1, l_{21}, \ldots, l_{2k}, \ldots, l_{2m_2})^T$ as the second, and third level variables respectively, (the first level is the evidence e), and allow

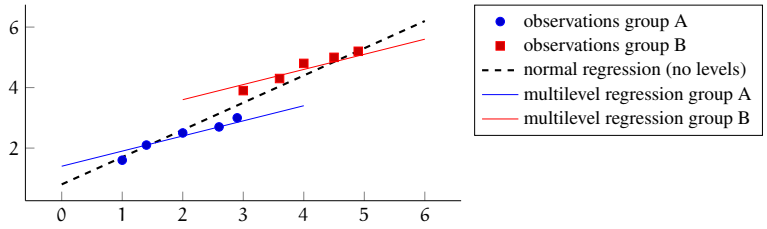


Figure 6.2: Multilevel regression, showing that the effect of x on y (the slope) is in fact lower than computed from normal regression. This effect is due to the fact that multilevel regression allows different a priori estimates (β_0) for each group.

dependency of β on the third level variables as well. The coefficient β is now a three-dimensional array consisting of components β_{ijk}. If the vector γ_{qs}, consisting of elements γ_{iqs}, represents the third level noise (with homogeneity of variances), the model becomes:

$$P(O_{qs} \mid e, l) \sim \mathcal{N}(\mu_{qs}, \sigma) \tag{38a}$$

$$\mu_{qs} = \left(\delta_q + ((\gamma_{qs} + \beta l_2)^T l_1\right)^T e \tag{38b}$$

where again the expectation $E[O_{qs} \mid e, l] = \mu_{qs}$.

This last model assumes the random outcome variable O to be normally distributed, but in case that O is dichotomous this no longer holds. In this case a specific transformation of the outcome variable, e.g., the logistic function, is assumed to be linear dependent of the explanatory variables. For logistic regression the transformation is given by:

$$\text{logit } E[O \mid e] = \log \frac{E[O \mid e]}{1 - E[O \mid e]},$$

and the logistic multilevel model therefore becomes:

$$\text{logit } E[O_{qs} \mid e, l] = \left(\delta_q + ((\gamma_{qs} + \beta l_2)^T l_1\right)^T e.$$

The conditional probability in case of logistic regression is defined as:

$$P(O_{qs} \mid e, l) \sim \text{Bernoulli}(p) \tag{39a}$$

$$\text{logit } p = \left(\delta_q + ((\gamma_{qs} + \beta l_2)^T l_1\right)^T e \tag{39b}$$

When actually doing the multilevel regression we might not want (or expect) an effect of certain higher levels variables on *all* lower level variables. In that case the corresponding component β_{ijk} is fixed to zero, i.e., it is omitted from the model.

Multilevel regression requires less parameters in comparison to standard regression, where the higher level variables are modelled as explanatory variables [83]. Parameters of multilevel regression models can be estimated using an iterative generalised least square (IGLS) method. IGLS is a least square method that estimates the parameters by alternating the optimising process between the fixed parameters (β_{ij}) and the

stochastic parameters (δ_{iq}) until convergence is reached. Goldstein [62] proved that this method is equivalent to the maximum likelihood estimation in standard regression, and improved it to restricted iterative generalised least square (RIGLS) which coincides with restricted maximum likelihood (REML) in Gaussian models [63]. Parameters for dichotomous outcomes are estimated with marginal and penalised quasi-likelihood (MQL/PQL) algorithms [18, 65]. Alternatively Markov chain Monte Carlo (MCMC) methods such as Gibbs sampling can be used [187]. Further information and comparison of Bayesian and likelihood-based methods for fitting multilevel models can be found in [21]. Note that a regression method always tries to fit the model on observed variables only, i.e., it does not consider unobserved variables. For more details about multilevel regression models one is referred to [83].

6.4 DEALING WITH MULTILEVEL DATA BY BAYESIAN NETWORKS

In multimorbidity it is of interest to study in which way diseases interact. For example, diseases D and D' might be unconditionally dependent of each other, i.e., $D \not\perp\!\!\!\perp_P D' \mid \varnothing$, but they could become independent if an environmental factor F is taken into account, $D \perp\!\!\!\perp_P D' \mid F$. This means that the factor F offers a complete explanation of the interaction between the disease D and D'. Moreover, when diseases are represented in a Bayesian network, the Markov blanket of a disease D corresponds to all factors, possibly other diseases, that are relevant for the prediction of this disease D.

In this section, we introduce the *multilevel Bayesian network* (MBN) formalism as a new model-based representation of multilevel data. As mentioned in the introduction, this combines the multilevel methodology, used in multilevel regression, with Bayesian networks, in such way that we are able to analyse interactions and probabilistic dependencies between multiple diseases, using patient data obtained from multiple sources, such as family practices.

6.4.1 *Basic ideas*

The advantage of a Bayesian network over regression models is that all variables are treated as uncertain, where in regression, including multilevel regression, only the outcome variable is treated as uncertain. If one is primarily interested in the interaction between all relevant variables, and not only in prediction of outcome, in the context of multiple diseases, this is a convenient way to model multiple diseases. Furthermore, as multilevel regression models can be seen as conditional probability distributions, they can be used as a factor in a Bayesian network (cf. Equation (1) in Chapter 2). In this section, we explore this relationship by varying the amount of structure in such models and compare this to the multilevel regression approach. However, the first challenge that must be met is the incorporation of multilevel methods in the Bayesian-network framework.

In multilevel regression, the random outcome variable O depends on the vectors of explanations of (non-random) variables, i.e., $e = (e_1, \ldots, e_n)$ and $l^j = (l^1_1, \ldots, l^j_{m_j})$, with $j = 1, \ldots, m$, (sub)groups q, and $m + 1$ different levels. For a Bayesian network

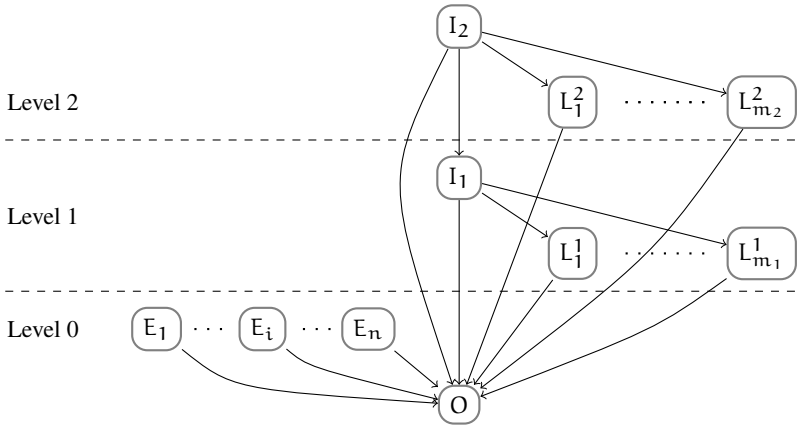


Figure 6.3: Bayesian network representation of a multilevel regression model.

approach, we model O as a conditional probability distribution given the set of parents $\{E_1, \ldots, E_n\} \cup \bigcup_{j=1}^{m} L^j$, with $L^j = \{L^j_1, \ldots, L^j_{m_j}\}$, and an *indicator variables* I_j, where $j = 1, \ldots, m$, that selects the group of objects at a certain level j. Figure 6.3 shows the corresponding Bayesian network with three levels, assuming no further dependence between variables. Clearly, this model is still too restrictive for most healthcare applications, as no structure is present between the explanatory variables and we have only one outcome variable of interest.

The idea of a multilevel Bayesian network is that the indicator variables I split the domain into different categories with a deterministic effect on the group variables L that is constant for a given category chosen by I. If not present, I variables can be constructed, e.g., by the Cartesian product in case of categorical L variables. However, multilevel analysis, and thus a multilevel Bayesian network, is typically designed for hierarchically structured data, and then the indicator I variables are part of the database definition.

Some of the explanatory variables are group-independent, though structure may exist between these variables. These variables correspond with the set of variables E in an MBN. Other variables, depend both on grouping and other variables at the same or higher levels. These variables correspond with the set of variables O in an MBN. The Bayesian network is constrained in the sense that no edges exist from a lower-level variable to a higher-level variable. This ensures that we keep the hierarchical structure present in multilevel regression methods. Because of the deterministic relations we are able to simplify the structure of the MBN using the following property.

Lemma 1. *Let X and Y be two random variables such that Y is deterministically dependent of X, i.e., there exists some function f such that* $Y = f(X)$. *Then, for all sets of random variables Z disjoint of X and Y it holds that* $Z \perp\!\!\!\perp_P Y \mid X$.

Proof. Take some arbitrary Z. If it is a discrete distribution, then it holds that:

$$P(Z \mid X) = \sum_Y P(Z, Y \mid X) = \sum_Y P(Z \mid X, Y) P(Y \mid X)$$

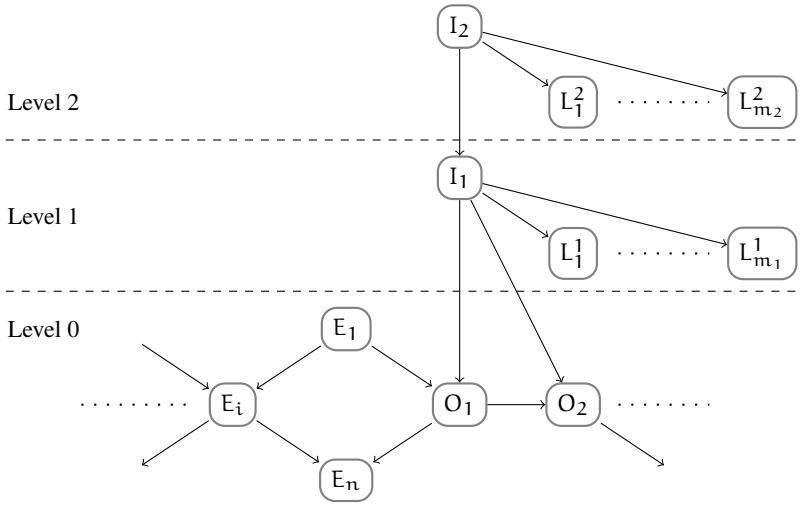


Figure 6.4: Multilevel Bayesian network with 3 levels and discrete variables.

By the relationship between X and Y, it holds $P(Y \mid X) = 1$ if $Y = f(X)$, and 0 otherwise, so it follows that:

$$P(Z \mid X) = P(Z \mid X, f(X)) = P(Z \mid X, Y)$$

Similarly, for continuous distributions, we have:

$$\begin{aligned} p(Z \mid X) &= \int p(Z, Y \mid X) dY = \int p(Z \mid X, Y) p(Y \mid X) dY \\ &= \int p(Z \mid X, Y) \delta(Y - f(X)) dY = p(Z \mid X, f(X)) \\ &= p(Z \mid X, Y) \end{aligned}$$

where δ is the Dirac delta function. □

We can apply this lemma to our initial MBN for two cases. Since $P(L_i^j \mid I_j)$ is deterministic, we obtain $O \perp\!\!\!\perp {}_P L_i^j \mid I_j$. The implication of this, is that no arcs exist between the group vertices in L and the outcome and explanatory vertices in $O \cup E$. Since the probability distribution $P(I_{j+1} \mid I_j)$ is deterministic too, we obtain $O \perp\!\!\!\perp {}_P I_j \mid I_1$ for all j. The implication of this is that within the indicator vertices I there are only arcs from I_{j+1} to I_j, for all j, and between the indicator vertices I and outcomes O there are only direct arc from I_1 to any O_i. These restrictions greatly simplify the structure of an MBN. When making predictions based on the parameters of the MBN, the indicator variables are mostly unknown. However, the structure still allows us to use the higher level variables to explain the outcome variable.

We now give a precise definition of MBNs. To shorten the definition, members of the various sets S are denoted by S_i, and $S \setminus \{S_i\}$ with $S - S_i$.

Definition 7. *A Bayesian network* $\mathcal{B} = (G, X_V, P)$ *is a* multilevel Bayesian network, *or MBN for short, if its set of vertices* V *is described by the tuple* (m, O, E, L, I), *with pairwise disjoint sets* $O, E, L, I \subseteq V_G$, *such that:*

- $m \in \mathbb{N}$ *denotes the* number of levels *of the MBN, where level 0 is called the* base level*;*
- O, *the set of* outcome variables, *is at base level such that if* $(V \to O_i) \in A_G$, *then* $V \in E \cup (O - O_i) \cup I$*;*
- E, *the set of* explanatory variables, *is at base level, such that if* $(V \to E_i) \in A_G$, *then* $V \in (E - E_i) \cup O$*;*
- $L = \{L^1, \ldots, L^m\}$, *where each* L^j *is a set of group variables at level* $j \geqslant 1$. *For group variable* L^j_i *it holds that*
  1. $(V \to L^j_i) \in A_G$ *implies that* $V = I_j$*;*
  2. $P(L^j_i \mid I_j)$ *is deterministic.*
- $I = \{I_1, \ldots, I_m\}$ *are* indicator variables, *such that* I_j *is the only parent of* I_{j-1} *in* G, *for all* $1 \leqslant j \leqslant m$, *and* $P(I_{j-1} \mid I_j)$ *is deterministic;*
- $X_V = \{X_v \mid v \in (I \cup E \cup O \cup L)\}$.

Figure 6.4 offers a graphical illustration of the definition. Note that within one MBN multiple diseases can be modelled as outcome variable. By lemma 1, the outcome variables O are independent of the level variables L given the value of the I variables. However, this does not imply that these variables are meaningless. Once the parameters of the MBN are learned, the level variables can be used to estimate the variance in the probability distribution of the outcome variables, without knowing the value of the indicator variables.

6.4.2 *Probability distributions for multilevel Bayesian networks*

Without taking into account the level variables, the probability of the outcome variables O conditioned on the explanatory variables E can be obtained by

$$\begin{aligned} P(O = o \mid E = e) &= f_{O|E}(o \mid e; \beta) \\ &= f_{O,E}(o, e; \beta) / \sum_o f_{O,E}(o, e; \beta), \end{aligned}$$

if O is discrete, and

$$\begin{aligned} P(O \leqslant o \mid E = e) &= \int_{-\infty}^{o} f_{O|E}(x \mid e; \beta) dx \\ &= \int_{-\infty}^{o} f_{O,E}(x, e; \beta) dx / \int_{-\infty}^{\infty} f_{O,E}(x, e; \beta) dx \end{aligned}$$

if O is continuous. The parameter β represents the parameters typically used for a specific distribution, e.g., $\beta = (\mu, \sigma)$ in case $f_{O,E}(o, e; \beta)$ is a Gaussian distribution with mean μ and variance σ.

In a multilevel Bayesian network the grouping variable splits the conditional probability distributions between an outcome variables and its explanatory variables into multiple (countable) distributions keeping them closely related, i.e., only the distribution type dependent parameters differ between groups. In case O is discrete we obtain $P(O = o \mid E = e, I = i) = f_{O|E}(o \mid e; \beta_i)$, and likewise, if O is continuous we obtain $P(O \leqslant o \mid E = e, I = i) = \int_{-\infty}^{o} f_{O|E}(x \mid e; \beta_i)dx$.

For example, in case O and $E = e$ are both discrete and O is binary with a Bernoulli distribution with parameter $\beta = p_{e,i}$, we obtain:

$$\begin{aligned} P(O = o \mid E = e, I = i) &= f_{O|E,I}(o \mid e, i; \beta) \\ &= \text{Bernoulli}(p_{e,i}) \\ &= \begin{cases} p_{e,i} & \text{if } O = o \\ 1 - p_{e,i} & \text{otherwise} \end{cases} \end{aligned}$$

In case O and E are both continuous and O follows a Gaussian distribution, we obtain the probability density function:

$$f_{O|E,I}(o \mid e, i; \beta) = \mathcal{N}(\mu_{e,i}, \sigma)$$

Just as in multilevel linear regression, a linear dependency between E and O can be obtained if $\mu_{e,i} = \beta_i e$, also for E being a discrete variable.

In case O is discrete and E is continuous a link function is used in multilevel regression, to keep the linearity in the model, of which the logistic function is the most popular one. The probability mass function for such a discrete variable with a continuous parent is:

$$f_{O|E,I}(o \mid e, i) = \frac{\exp\left(\beta^i_{o0} + \beta^i_{o1} e\right)}{\sum_o \exp\left(\beta^i_{o0} + \beta^i_{o1} e\right)}$$

For binary outcome variables this reduces to:

$$f_{O|E,I}(o \mid e, i) = \left[1 + \exp\left(\beta^i_0 + \beta^i_1 e\right)\right]^{-1}$$

6.5 EXPERIMENTAL METHODOLOGY

In the previous section, the basic ingredients of multilevel Bayesian networks were outlined. In this section, we take the step in making the technique practically useful. At the end of this section, we demonstrate that the methodology works by using synthetic data. In the next section, the same is done, but then for a dataset obtained from a public health registry containing patient data from general practices.

6.5.1 *Parameter learning*

Because we have incorporated the multilevel regression model as factors in the model, we can make use of multilevel regression to estimate the outcome variables. This has

the advantage that we exploit the correlation between different groups (if it exists) and therefore requires less data per group than a standard Bayesian network learning algorithm needs for parameter learning per group. For multilevel-level logistic regression models, it is recommended to use a minimum group size of 50 with at least 50 groups to produce valid estimates [134]. An exact inference algorithm for parameter estimation in networks with discrete children of continuous parents is proposed in [109]. Compared to multilevel regression models, it is also possible to use a Bayesian approach for learning the parameters [195] and, therefore, include even more domain knowledge to the model.

6.5.2 *Model validation*

Possible criteria to validate the model parameters are the Akaike information criterion (AIC) [2], the Bayesian information criterion (BIC) [183], and the deviance information criterion (DIC) [196]. The AIC and BIC are widely accepted decision criteria, but computationally expensive when dealing with large amounts of data and MCMC methods. This problem is overcome using the DIC, which calculates deviance residuals, that sum up to the deviance statistic, along with the MCMC process. Unfortunately, in disease mapping, DIC is in favour of overparameterised models, especially when using large datasets [160].

Alternatively, an approximation method proposed by [128] can be used, which works very well for large data sets in an MCMC setting. It uses replication of the stochastic parameters and the outcome variables for a specified part of the data along with the MCMC simulation based on the remaining part of the data. The replicate outcome variables can then be compared to the real outcomes, allowing us to assess the predictability of the model.

Although computationally expensive as well, standard cross validation (e.g., k-fold cross validation) is a robust method to validate regression and Bayesian models [158], and receiver operating characteristic (ROC) analysis can be used to validate accuracy and precision of the model parameters. Recently, a new measure was introduced, the net reclassification improvement (NRI), offering additional incremental information compared to the area under the curve (AUC) within an ROC analysis [155], which provides more insight into risk prediction.

6.5.3 *Structure learning*

In order to build the structure between variables, we can make use of two approaches. We can either model the structure manually based on existing medical knowledge or learn the structure from data. Structure learning of Bayesian networks offers a suitable method to learn these dependencies. The constraints imposed by the multilevel Bayesian network can be captured by blacklisting and whitelisting edges, which can be incorporated into a wide range of structure learning algorithms (see, e.g., [185]). For example, the necessary edges between I_1 and all variables $O_i \in O$ are whitelisted, whereas edges from a lower level to a higher level are all blacklisted.

A systematic approach to identify statistically significant edges in a network, has been developed by Friedman et al. using bootstrap resampling and model averaging [56]. The empirical probability of an edge, defined as the fraction of occurrences in the networks learned from bootstrapped samples, are known as edge intensities (or strengths), and can be interpreted as the degree of confidence that the edge is present in the true network structure describing the true dependence structure of the original data. Scutari et al. propose a statistically motivated estimator for the confidence threshold minimizing a specific norm between the cumulative distribution function of the observed confidence levels and the cumulative distribution function of the confidence levels of the unknown true network [186]. Classical norms are the rectilinear distance, denoted as the L_1 norm, and the euclidean distance, denoted as the L_2 norm [95].

6.5.4 *Artificial multimorbidity example with synthetic data*

Suppose we have the variables D_1, D_2, and D_3 that model whether the diseases D_1, D_2, and D_3 are present, a genetic variable G, and two demographic vertices L_1 and L_2 that model certain environmental conditions. Furthermore, let the demographics be variables obtained from higher levels in a hierarchically structured dataset, i.e., L_1 and L_2 are level-2 and level-3 variables respectively. For example, the variables D_1, D_2, and D_3 could represent diseases like *diabetes*, *retinopathy*, and *hypertension*. The variable G could represent *gender*, or a specific *gene*, and the grouping variable I_1 could represent a division in *practices* with *type* as L_1, and I_2 a division in *area* with *urbanity* as L_2.

There are fifty *practices* ($I_1 \in \{1,\dots,50\}$) and five *areas* ($I_2 \in \{1,2,3,4,5\}$). The variable *type* (L_1) has 5 possible values and the variable *urbanity* (L_2) is binary. The deterministic relations between them are:

$$I_2 = \begin{cases} 1 & \text{if } I_1 \in \{1,\dots,10\} \\ 2 & \text{if } I_1 \in \{11,\dots,20\} \\ 3 & \text{if } I_1 \in \{21,\dots,30\} \\ 4 & \text{if } I_1 \in \{31,\dots,40\} \\ 5 & \text{if } I_1 \in \{41,\dots,50\} \end{cases}$$

and

$$U_1 = \begin{cases} 1 & \text{if } I_1 \bmod 10 \in \{0,1\} \\ 2 & \text{if } I_1 \bmod 10 \in \{2,3\} \\ 3 & \text{if } I_1 \bmod 10 \in \{4,5\} \\ 4 & \text{if } I_1 \bmod 10 \in \{6,7\} \\ 5 & \text{if } I_1 \bmod 10 \in \{8,9\} \end{cases}$$

and

$$U_2 = \begin{cases} 0 & \text{if } I_2 \in \{1,2\} \\ 1 & \text{if } I_2 \in \{3,4,5\} \end{cases}$$

We sampled 10,000 patients uniformly over the fifty practices and determined its respective values for the other higher level variables. The binary variable G is binomially sampled with a probability of 0.50. The diseases D_1, D_2 and D_3 are sampled as follows.

$$\begin{aligned} D_1 &= \text{Binomial}(0.50 + 0.01G + \mathcal{N}(\mu_q, \sigma_q)) \\ D_2 &= \text{Binomial}(0.20 + \mathcal{N}(\mu_s, \sigma_s)) \\ D_3 &= \text{Binomial}(0.20 + 0.1D1 + 0.2D2 + 0.2D1D2 + \mathcal{N}(\mu_{qs}, \sigma_{qs})) \end{aligned}$$

With $q = 1, \ldots, 50$, and $s = 1, \ldots, 5$, corresponding to the number of *practices* and *areas*. The distributions $\mathcal{N}(\mu_q, \sigma_q)$ and $\mathcal{N}(\mu_{qs}, \sigma_{qs})$ are randomly sampled from a $\mathcal{N}(0, 0.1)$ distribution, μ_s is 0.25, 0.30, 0.35, 0.40, and 0.45, for $s = 1, \ldots, 5$ respectively, and $\sigma_s = 0.01$.

Applying multilevel regression, if we, for example, only allow an influence of the level-2 and level-3 variables on the intercept and the regression coefficient of the explanatory variable D_1, the multilevel regression model becomes:

$$P(D_{3qs} \mid d_1, d_2, g, l_1, l_2) \sim \text{Bernoulli}(p) \tag{40.1a}$$

$$\text{logit } p = \beta_{0qs} + \beta_{1qs} d_1 + \beta_2 d_2 + \beta_3 g \tag{40.1b}$$

$$\beta_{0qs} = \beta_{00s} + \beta_{01s} l_1 + \delta_{0q} \tag{40.2a}$$

$$\beta_{1qs} = \beta_{10s} + \beta_{11s} l_1 + \delta_{1q} \tag{40.2b}$$

$$\beta_{00s} = \beta_{000} + \beta_{001} l_2 + \gamma_{00s} \tag{40.3a}$$

$$\beta_{01s} = \beta_{010} + \beta_{011} l_2 + \gamma_{01s} \tag{40.3b}$$

$$\beta_{10s} = \beta_{100} + \beta_{101} l_2 + \gamma_{10s} \tag{40.3c}$$

$$\beta_{11s} = \beta_{110} + \beta_{111} l_2 + \gamma_{11s} \tag{40.3d}$$

With $\delta_{iq} \sim N(0, \sigma_{iq})$ and $\gamma_{iqs} \sim N(0, \sigma_{iqs})$. Substituting the equations of 40.3 into 40.2, and the equations of 40.2 into 40.1, equation 40.1 becomes:

$$\begin{aligned} \text{logit } p &= (\beta_{000} + \beta_{001} l_2 + \gamma_{00s} + (\beta_{011} l_2 + \gamma_{01s} + \beta_{010}) l_1 + \delta_{0q}) \\ &+ (\beta_{100} + \beta_{101} l_2 + \gamma_{10s} + (\beta_{111} l_2 + \gamma_{11s} + \beta_{110}) l_1 + \delta_{1q}) d_1 \\ &+ \beta_2 d_2 + \beta_3 g \end{aligned}$$

Since L_1 and L_2 are discrete variables, and have the same value within a group, we can rewrite this into:

$$\begin{aligned} \text{logit } p &= (\beta'_0 + \beta'_{0s} + \gamma'_{0s} + \beta'_{0qs} + \gamma'_{0qs} + \beta'_{0q} + \delta_{0q}) \\ &+ (\beta'_1 + \beta'_{1s} + \gamma'_{1s} + \beta'_{1qs} + \gamma'_{1qs} + \beta'_{1q} + \delta_{1q}) d_1 \\ &+ \beta_2 d_2 + \beta_3 g \end{aligned}$$

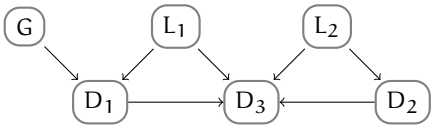


Figure 6.5: Bayesian network representing probabilistic dependencies between certain diseases (D_1,D_2,D_3), a genetic variable G, and some demographics (L_1,L_2).

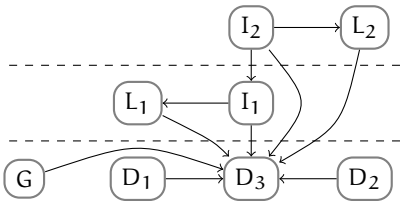


Figure 6.6: MBN representing multilevel regression of the example in Figure 6.5.

Now, assume that using structure learning (without using the indicator variables) it is observed that $\pi(G) = \pi(L_1) = \pi(L_2) = \varnothing$, $\pi(D_1) = \{G, L_1\}$, $\pi(D_2) = \{L_2\}$, and $\pi(D_3) = \{D_1, D_2, L_1, L_2\}$. Figure 6.5 then shows the corresponding Bayesian network and the joint distribution $P(V)$ is given by

$$P(D_3 \mid D_1, D_2, L_1, L_2)P(D_1 \mid L_1, G)P(D_2 \mid L_2)P(L_1)P(L_2)P(G)$$

To predict whether a disease D_3 is present given that L_1, L_2 and G are known, we have by the definition of a Bayesian network and standard probability theory:

$$P(D_3 \mid L_1, L_2, G) = \sum_{D_1, D_2} P(D_3 \mid D_1, D_2, L_1, L_2)P(D_1 \mid L_1, G)P(D_2 \mid L_2)$$

Since the Markov blanket of D_3 is $\{D_1, D_2, L_1, L_2\}$, any information about the genetic variation of a person is irrelevant, i.e., since $D_3 \perp\!\!\!\perp {}_P G \mid D_1$ we obtain: $P(D_3 \mid D_1, D_2, L_1, L_2, G) = P(D_3 \mid D_1, D_2, L_1, L_2)$.

Applying the MBN techniques, Figure 6.6 and 6.7 show the corresponding MBN representations. One can see that in the multilevel regression network (Figure 6.6) only D_3 is modelled as an outcome variable of interest, as where in the structured model (Figure 6.7) D_1 and D_2 are modelled as outcome variables as well (still being an explanatory variables of D_3). As a consequence of Definition 7, the disease variables in Figure 6.7 do not have edges from I_2 and L_i directed towards themselves.

The comparison between the multilevel regression technique and the structured multilevel Bayesian network is outlined in table 6.1, showing the probability of disease D_3 in the presence of L_1, D_1 and D_2. Parameters of the multilevel regression model are obtained with the MLWin software, in which the algorithms described at the end of section 6.3 are implemented [66]. Parameters of the MBN are learned using the *bnlearn* package [185] in the statistical software R.

Using AIC and BIC, the most accurate multilevel logistic regression model allows random intercepts and random slopes on D_1 for each entry of L1. Although the probabilities derived from the MBN are closer to the true probabilities, the area under the

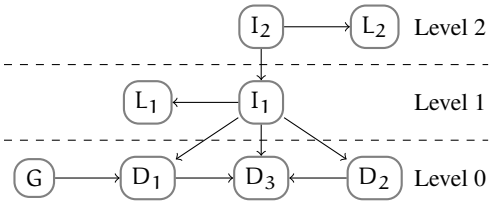


Figure 6.7: Structured MBN representation of the example in Figure 6.5.

	$L_1 = 1$	$L_1 = 2$	$L_1 = 3$	$L_1 = 4$	$L_1 = 5$
$D_1 = 0, D_2 = 0$	0.150	0.175	0.200	0.225	0.250
$D_1 = 0, D_2 = 1$	0.250	0.275	0.300	0.325	0.350
$D_1 = 1, D_2 = 0$	0.350	0.375	0.400	0.425	0.450
$D_1 = 1, D_2 = 1$	0.650	0.675	0.700	0.725	0.750

(a) True probability distributions in the test set

	$L_1 = 1$	$L_1 = 2$	$L_1 = 3$	$L_1 = 4$	$L_1 = 5$
$D_1 = 0, D_2 = 0$	0.135	0.137	0.171	0.187	0.198
$D_1 = 0, D_2 = 1$	0.279	0.281	0.319	0.335	0.346
$D_1 = 1, D_2 = 0$	0.410	0.414	0.479	0.505	0.523
$D_1 = 1, D_2 = 1$	0.632	0.635	0.676	0.692	0.702

(b) Multilevel logistic regression (Equation 40)

	$L_1 = 1$	$L_1 = 2$	$L_1 = 3$	$L_1 = 4$	$L_1 = 5$
$D_1 = 0, D_2 = 0$	0.164	0.148	0.189	0.228	0.218
$D_1 = 0, D_2 = 1$	0.271	0.242	0.286	0.304	0.331
$D_1 = 1, D_2 = 0$	0.330	0.398	0.442	0.453	0.466
$D_1 = 1, D_2 = 1$	0.653	0.680	0.747	0.735	0.749

(c) Structured multilevel Bayesian network (Figure 6.7)

Table 6.1: Probability estimations of D_3 conditioned on D_1, D_2 and L_1.

curves (AUCs) within an ROC analysis are close together, i.e., 0.725 and 0.712 for the MBN and multilevel regression respectively. In the multilevel regression all variables are used for prediction, whereas for the MBN only the variables of the Markov blanket are used for prediction.

The net reclassification improvement is in favour of the MBN, i.e., the NRI is 0.2144 ($p < 0.001$). Thus, on average the MBN is significantly better then the multilevel regression approach in this synthetic example. This due to the fact that an MBN is able to give an exact solution with respect to a dependency structure between variables and its observations. Multilevel regression does not have these dependency constraints, which possibly favours overfitting the model.

6.6 MODELLING INTER-PRACTICE VARIATION IN MULTIMORBIDITY

Normally, in scientific research, one would investigate diseases separately, resulting in different predictive values of variables shared by both diseases. As an exaple, multilevel regression analysis was recently used by Nielen et al. to investigate the influence of particular family practice variables on hypertension and diabetes mellitus separately, revealing an inter-practice variance in predictability [140]. However, since interactions could have an additive effect on prevalence, this yields no insight into the predictive value in case both diseases are present. In fact, we need an extra regression model on the combined diagnosis of hypertension and diabetes together to be able to draw such conclusions.

In this chapter, we will use the research of Nielen et al. as starting point. Firstly, we compare the parameter estimations of an unstructured MBN with multilevel regression. Secondly, we compare the predictive power of a structured MBN with multilevel regression.

6.6.1 *Description of the models*

To evaluate if the parameter estimations of an MBN are comparable with a multilevel regression we analysed models for both diabetes mellitus and heart failure. Nielen et al. analysed hypertension instead of heart failure. However, besides the validation of parameter estimations, we also want to investigate the predictive power for diseases that have a different onset during lifetime. Heart failure is known to be associated with diabetes mellitus and hypertension [78], and its risk management involves almost the same variables [223]. Since the onset of hypertension and diabetes mellitus is typically earlier in the patient's life than the onset of heart failure it is of interest if the finally structured MBN follows this order.

We used five models for the analysis. The first two models are the multilevel regression models for predicting either diabetes mellitus (model MLR-DM) or heart failure (model MLR-HF) using data which is grouped by practice, where the urbanity of the practice is modelled as as higher level variable. The next two models (MBN-DM and MBN-HF) are the corresponding unstructured MBNs for the first two models, assuming that no further dependencies between variables exist (cf. Figure 6.3), and that the urbanity is independent of the disease, given the practice (cf. Property 1). Finally, we

consider a structured model (MBN-STR) which contains both diseases as well as structure between the outcome and explanatory variables, which we call intra-level structure.

All five models used practice and urbanity as higher level variables. Since the practices use different types of information systems, one could expect that this might also be of influence on the predictions. To model this, a second level grouping variable (the used information system) can be incorporated on top of the first level grouping variable (practice). However, it appeared that there was no significant benefit from doing so. Therefore, this extra variable was omitted from further analysis.

6.6.2 *Research problem and data*

The patient data was routinely collected by the Netherlands information network of general practice (LINH); see Section 3.4 for details. In the analysis here, patients under 25 years were excluded, because of their low probability of multimorbidity. Practices which recorded during less than six month were also excluded from statistical analysis. Eventually, we used data of 218,333 patients from 82 Dutch general practices, meaning an average number of patients around 2650 per practice. Morbidity data were derived from diagnoses, using the international classification of primary care (ICPC) and anatomical therapeutic chemical (ATC) codes.

6.6.3 *Unstructured MBNs compared to multilevel regression*

For both the multilevel regression models MLR-DM and MLR-HF we estimated the parameters using MLWin [66]. For the models MBN-DM and MBN-HF we used MCMC simulation, available in the WinBUGS software [195]. All non-group variables were discretised and modelled using a Bernoulli distribution. Parameter estimates using a 10-fold cross validation are presented in Table 6.2. As expected, the results of the unstructured MBN models are similar to the results obtained by multilevel regression, showing that multilevel Bayesian networks are a valid alternative method for multilevel analysis.

6.6.4 *Composition of the structured MBN*

The structure of the MBN-STR model was learned using the *bnlearn* package [185] that is part of the statistical software R. It provides various methods for structure learning. We have restricted the search of Bayesian networks to those that satisfy the multilevel structure by using white- and blacklists. See Figure 6.10 for the resulting Bayesian network structure. Note that indeed there is only a dependency between consecutive levels, and that this is solely through the grouping variables. Furthermore, it appeared that only a subset of the disease variables depends on the practice variable; diabetes mellitus is amongst them, whereas heart failure is not. So technically, diabetes mellitus is an outcome variable and heart failure is an explanatory variable within the definition of an MBN. However, since all variable can be treated as uncertain we can still use the model to make predictions for heart failure.

	Diabetes mellitus		Heart failure	
Model	MLR-DM	MBN-DM	MLR-HF	MBN-HF
Age	1.029	1.028	1.106	1.106
Gender (ref = male)	0.914	0.915	0.823	0.815
Overweight/obesity	1.725	1.671	1.689	1.600
Diabetes mellitus	-	-	1.256	1.260
Lipid disorder	6.437	6.392	1.172	1.183
Hypertension	5.675	5.800	2.071	2.067
Peripheral artery disease	0.954	0.949	1.619	1.530
Heart failure	1.132	1.194	-	-
Retinopathy	9.253	9.669	1.310	1.104
Angina pectoris	0.679	0.665	2.214	2.184
Stroke / CVA	0.770	0.766	1.388	1.397
Renal disease	1.176	1.200	1.878	1.881
Cardiovascular symptoms	0.848	0.850	2.596	2.636
Urbanity (ref=urban)				
urban	1.000	1.000	1.000	1.000
strongly urban	1.261	1.275	1.145	1.158
modestly urban	1.477	1.490	1.181	1.192
little urban	1.436	1.408	1.422	1.456
not urban	1.474	1.259	1.335	1.318

Table 6.2: Parameter estimations of explanatory (parent) variables, represented as odds ratios, using cross validation in a multilevel analysis for diabetes mellitus and heart failure (MLR=multilevel regression, MBN=multilevel Bayesian network, DM=diabetes mellitus, HF=heart failure).

Some of the directions of edges is opposite to what the domain experts would expect, e.g., angina pectoris is pointing towards peripheral artery disease (PAD), but in reality this is seen as a comorbidity due to atherosclerosis, which itself is not present in the model. Therefore, we also incorporated some domain knowledge [223, 78] into the model and allowed a geriatric specialist and two physicians to validate the model. Removed edges are: *angina pectoris* $\rightarrow$ *PAD*, *angina pectoris* $\rightarrow$ *renal disease*, *heart failure* $\rightarrow$ *PAD*, and *practice* $\rightarrow$ *cardiovascular symptoms*. The edge *heart failure* $\rightarrow$ *renal disease* is reversed. The final model is showed in Figure 6.8, along with the prior probability distributions for patients aged over 65 years. However, these results are of a preliminary nature, and we did not study the validity of the structured model further.

Using bootstrapped samples to validate the strengths of the edges, most edges shown in the network of Figure 6.8 appear in more than 95% of the networks learned from the samples. The only edges with a percentage lower than 95% is *renal disease* $\rightarrow$ *heart failure* (0.73%). Most of the edges not present in the originally learned structure have an appearance close to 0%.

In this model the prevalence rate of diagnosed diabetes mellitus in practices varies between 0.008 and 0.135, with mean 0.077 and standard deviation 0.025. The prevalence of heart failure varies between 0.001 and 0.059, with mean 0.019 and standard deviation 0.011. Figure 6.9 shows the same model as in Figure 6.8, but now conditioned on hypertension and diabetes, i.e., both diseases are present. In this case probabilities are more or less doubled (or tripled in case of lipid disorder), indicating the population of elderly patients with both hypertension and diabetes have twice the chance of getting an additional cardiovascular disease when compared to the general elderly population. For this population, i.e., diabetics with hypertension, the prevalence of heart failure varies between 0.001 and 0.230, with mean 0.086 and standard deviation 0.049.

Finally, the conditional probability distribution of a disease variable can be used to uncover interactions between diseases. If we calculate the probability of angina pectoris (ap) in the presence of both hypertension (ht) and lipid disorders (ld), we obtain: $\mathrm{P}(\mathtt{ap} \mid \mathtt{ht}, \mathtt{ld}) \approx 16\%$. It turns out that this is much higher than one can expect from the other probabilities: $\mathrm{P}(\mathtt{ap} \mid \mathtt{ht}, \overline{\mathtt{dl}}) \approx 7\%$, $\mathrm{P}(\mathtt{ap} \mid \overline{\mathtt{ht}}, \mathtt{ld}) \approx 5\%$ and $\mathrm{P}(\mathtt{ap} \mid \overline{\mathtt{ht}}, \overline{\mathtt{ld}}) \approx 1\%$. We can do this exercise for an arbitrary disease and (a subset of) its parents in the MBN structure. For example, when looking at heart failure (hf), there is an interaction between hypertension and diabetes mellitus (dm): $\mathrm{P}(\mathtt{hf} \mid \mathtt{ht}, \mathtt{dm}) \approx 9\%$, $\mathrm{P}(\mathtt{hf} \mid \mathtt{ht}, \overline{\mathtt{dm}}) \approx 5\%$, $\mathrm{P}(\mathtt{hf} \mid \overline{\mathtt{ht}}, \mathtt{dm}) < 1\%$ and $\mathrm{P}(\mathtt{hf} \mid \overline{\mathtt{ht}}, \overline{\mathtt{dm}}) < 1\%$; which suggest that the effect of diabetes on heart failure is only of clinical significance in the presence of hypertension.

6.6.5 *Comparison of the structured MBN with multilevel regression*

Besides the estimation of odds, a more practical question is how well the model can be used for prediction. For this, we compared the predictive performance of the MBN-STR model to multilevel regression analysis for single diseases, i.e., the models MLR-DM and MLR-HF.

For the multilevel regression method, we used all the predictors, while for the MBN-STR model, we can restrict ourselves to the Markov blankets (cf. Section 3.1) of the diseases and higher level variables where necessary. For diabetes mellitus, the MB

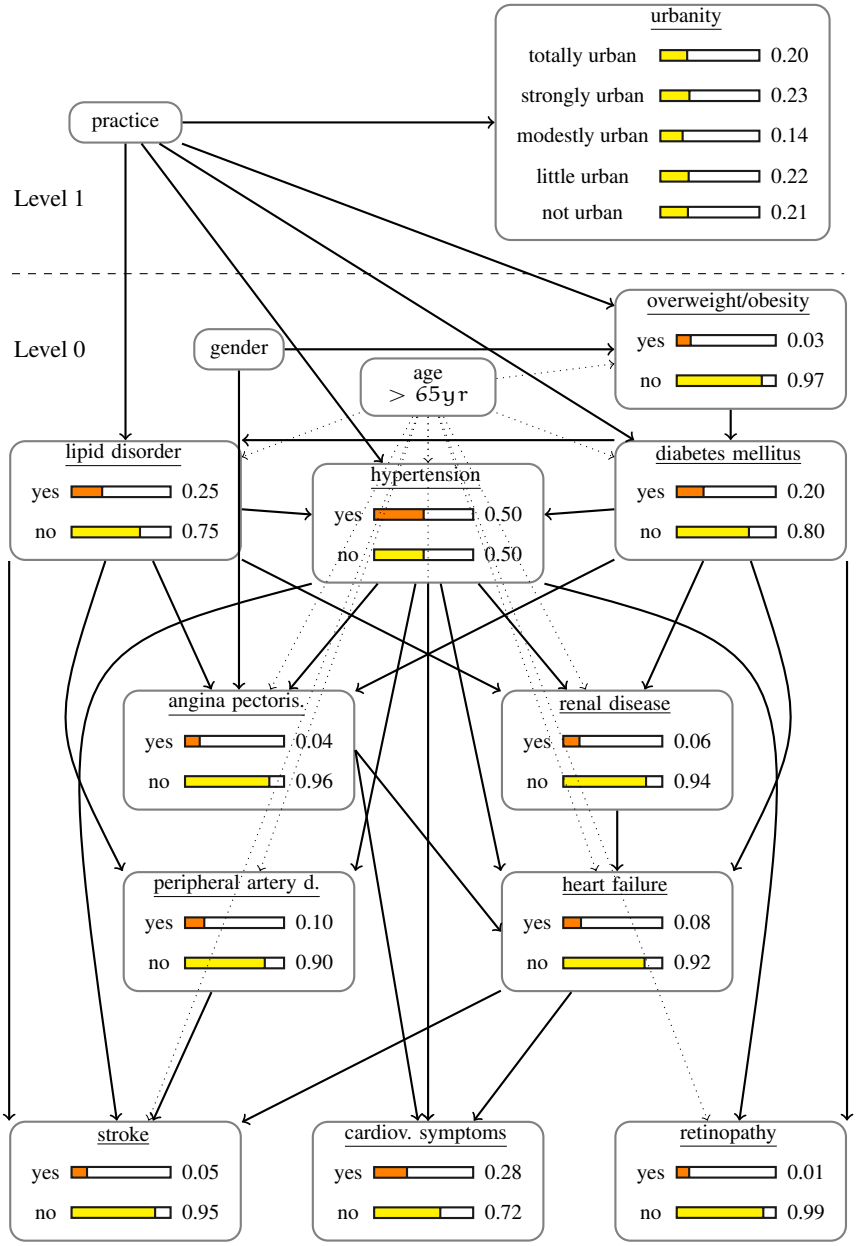


Figure 6.8: Structured MBN with prior probability distributions for patients aged > 65 years, using domain knowledge (expert opinions / evidence from other research) of cardiovascular diseases and diabetes mellitus in family practices. The dotted arcs are arcs from 'age' and 'gender' in order to make the model more readable.

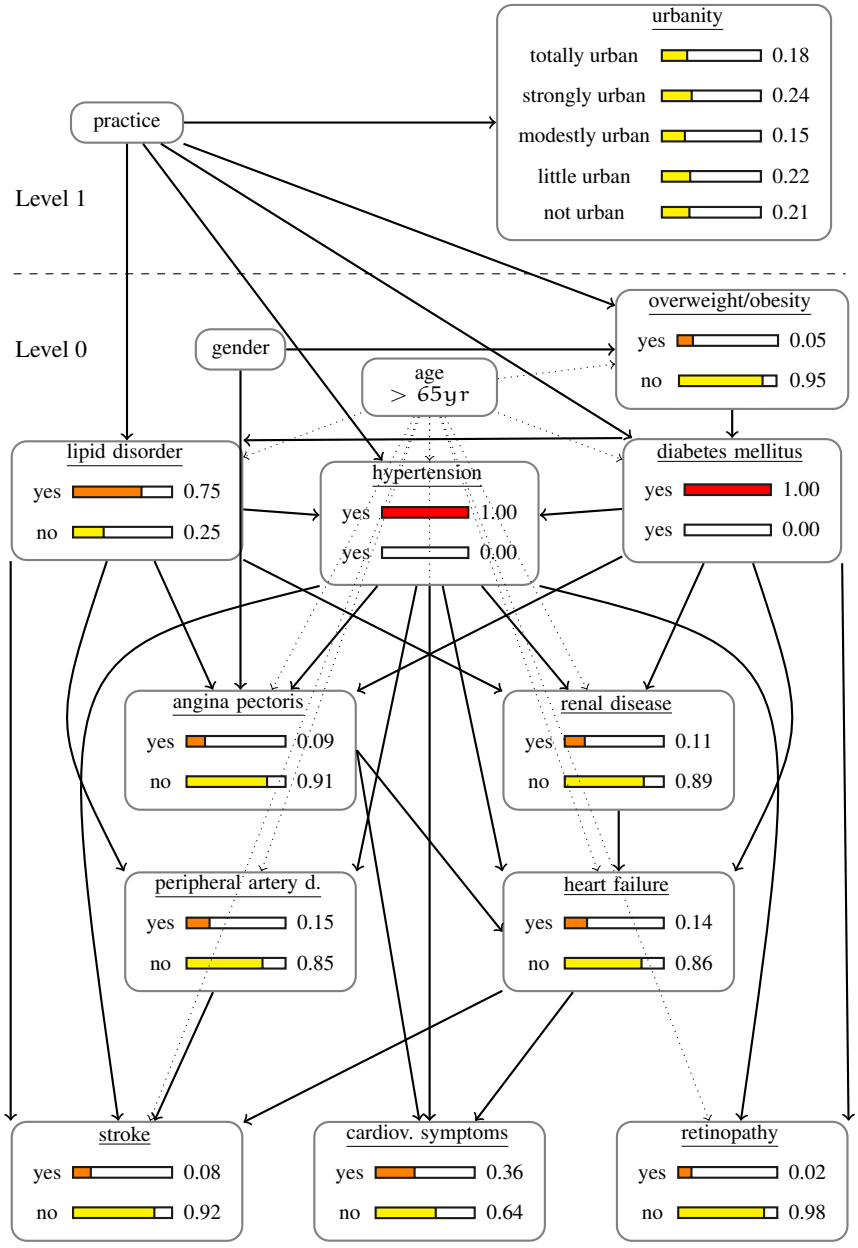


Figure 6.9: Structured MBN (cf. Figure 6.8) with posterior probability distributions for patients with both hypertension and diabetes (aged > 65 years).

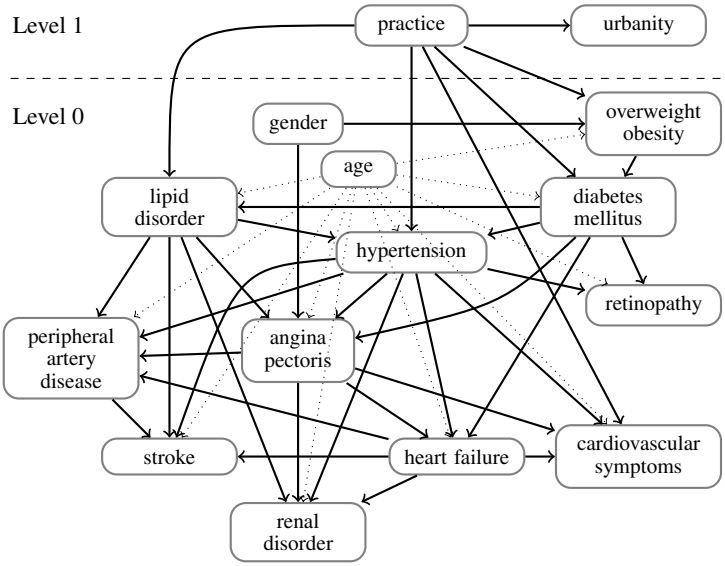


Figure 6.10: Structure learning without any domain knowledge of cardiovascular diseases and diabetes mellitus in family practices. The dotted arcs are arcs from 'age' in order to make the model more readable.

consists of practice, age, gender, obesity, lipid disorder, hypertension, heart failure, retinopathy, and renal disorder. However, making predictions in a multilevel model we treat the indicators, i.e., the practice, as uncertain, and instead we have to use the urbanity for prediction as well. The MB of heart failure on the other hand consists of age, gender, lipid disorder, diabetes mellitus, hypertension, peripheral artery disease, angina pectoris, stroke, renal disorder, and cardiovascular symptoms. For heart failure no higher level variables are needed for prediction when the diseases that vary along such variables are known, e.g., obesity, hypertension, and diabetes.

To measure the accuracy of the predictions we performed an ROC analysis (see Figure 6.11). When comparing the AUC between multilevel regression and the MBN-STR model, the ones for the MBN-STR model are slightly better with a difference of approximately 1%. For the MBN-STR they are approximately 0.90 and 0.84 for diabetes mellitus and heart failure respectively. For the MLR-DM it is 0.89 and for the MLR-HF it is 0.83. When performing a net reclassification improvement analysis for the MBN-STR model compared to the multilevel regression models MLR-DM and MLR-HF, the NRI is significantly positive in both cases, i.e., the NRI is 0.723 ($p < 0.001$) for diabetes and 0.075 ($p < 0.01$) for heart failure.

6.7 DISCUSSION

In this chapter, we have presented a new type of multilevel modelling, and applied this to healthcare data of general practices. As we have discussed, such data often contain a hierarchical structure, which can be modelled by using different levels of data. Since

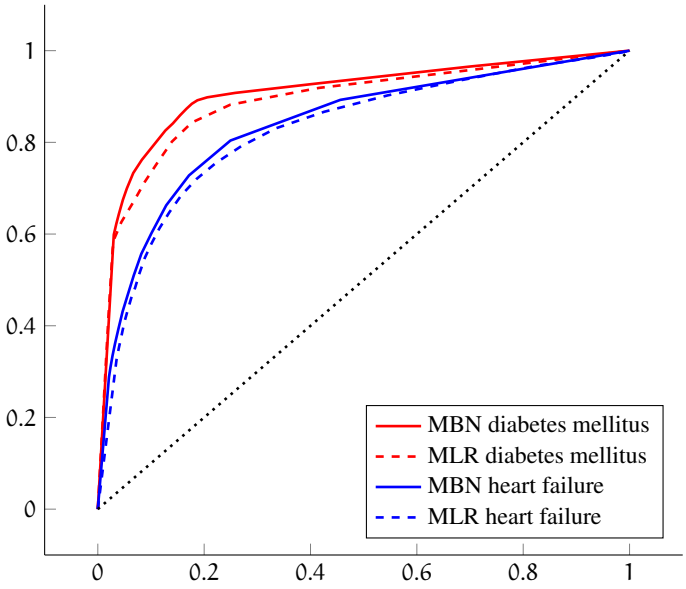


Figure 6.11: ROC analysis of a structured multilevel Bayesian network (MBN) and multilevel regression (MLR) for diabetes mellitus and heart failure.

traditional multilevel regression equations only include one outcome variable each time, which is unpractical in the context of multiple diseases, we combined Bayesian networks with multilevel analysis yielding multilevel Bayesian networks. Multilevel Bayesian networks also allow representing uncertainty of all disease variables concerned into one model.

Furthermore, we can add intra-level structures between variables giving extra insight into probabilistic dependencies and interactions. Moreover, domain knowledge can be incorporated, e.g., edges between pathophysiology and its corresponding lab results are always pointing to the latter, making the model more easy to interpret. Such domain knowledge can be used to restrict the search space when learning the structure of a Bayesian network from data.

We have shown that multilevel Bayesian networks have at least the expressive power of traditional multilevel regression methods. Using synthetic data and a real-world application of MBNs with clinical patient data from family practices, we demonstrated the empirical equivalence of a traditional multilevel regression model to an unstructured MBN. Furthermore, structured MBNs provide insight into the relationship between multiple diseases and allow for studying multiple diseases at the same time, avoiding the redundancy of regression methods (when used to analyse multiple disease for the same variable set).

Although it was not our main aim to develop a better classifier, the predictive power of a structured MBN is just as good as multilevel regression equations, despite a reduced number of predictors, as defined by the Markov blanket. Both in the synthetic example and the real-life applications of diabetes mellitus and heart failure, there is a

small improvement in the AUC and a significantly positive NRI. Bootstrapped samples showed that the strength of the edges between disease variables in the network representation of diabetes mellitus and heart failure is mostly close to 100%, meaning we can be confident about the found structure.

Using the learned MBN we are able to condition on certain disease variables, e.g., when conditioning on hypertension and diabetes, the MBN reveals that chances on obtaining another cardiovascular disease, such as heart failure, is more or less doubled. This 'personalisation' of the network could be seen as a step forward to personalised clinical guidelines, as mentioned in the introduction, making the MBN a promising tool in the new domain of multimorbidity. Further research will focus on the application of the MBN framework to relevant clinical questions within public health and the related multimorbidity issues.

Finally, since from the data available it will not be possible to construct a full causal model, it is important to make use of expert background knowledge. Besides placing restrictions on existing variables, one might also introduce variables that are missing from the data, this way adding crucial explanatory power. This is possible in BNs, and thus MBNs can also use the same expertise to quantify the probabilistic relationships involving these missing variables even though no data exists for them. As an example, atherosclerosis may be added to the model, and, using the method proposed in [209], this variable may capture important combinations of observations, e.g., peripheral artery disease along with a cardiac disease such as angina pectoris. This may improve the predictive performance of these models further.

7

MULTILEVEL TEMPORAL BAYESIAN NETWORKS CAN MODEL LONGITUDINAL CHANGE IN MULTIMORBIDITY

ABSTRACT

Whereas the course of a single disease can be studied well using traditional epidemiological methods, these methods cannot capture the complex joint evolutionary course of multiple disorders. In this study, multilevel temporal Bayesian networks were adopted to study the course of multimorbidity in the expectation that this would yield new clinical insight. Clinical data of patients were extracted from ninety general practice registries in the Netherlands. One and half million patient years were used for analysis. The simultaneous progression of six chronic cardiovascular conditions was investigated, correcting for both patient and practice-related variables. Cumulative incidence rates of one or more new morbidities rapidly increase with the number of morbidities present at baseline, ranging up to 47% and 76% for three and five-year follow-up respectively. Hypertension and lipid disorders, as health risk factors, increase the cumulative incidence rates for both individual and multiple disorders. Moreover, in their presence, the *observed* cumulative incidence rates of combinations of cardiovascular disorders, i.e., multimorbidity, differ significantly from the *expected* rates. There are clear synergies between health risks and chronic diseases when multimorbidity within a patient progresses over time. The method used here supports a more comprehensive analysis of such synergies compared to what can be obtained by traditional statistics.

7.1 INTRODUCTION

The epidemiology of multiple chronic diseases present at the same time is referred to as *comorbidity* or *multimorbidity*. Whereas comorbidity is usually defined in relationship to a specific index condition, as in the seminal definition of Feinstein [50], the term 'multimorbidity' has been introduced to refer to any co-occurrence of two or more medical, especially chronic, conditions within a person at the same time [206, 208]. In Chapter 3, we showed how multimorbidity can be simply measured by computing various basic statistics: the number of chronic disorders per patient, corrected for age, gender, and socio-economical demographics [207, 224, 53, 201], odds that describe the ratio between observed and expected prevalence rates for specific disease combinations [177, 68, 19, 215], and disease clusters using principal component analysis [81, 179, 124, 92].

Key findings:

- The urbanisation level of a general practice is associated with the cumulative incidence of chronic cardiovascular conditions, in particular those with a high prevalence, i.e., obesity, hypertension, dyslipidemia, diabetes mellitus, and ischaemic heart disease.
- The overall multimorbidity rate of chronic cardiovascular (related) disorders rapidly increases when multimorbidity is already present at baseline.
- When multimorbidity progresses over time, certain disease combinations develop more quickly than what can be expected from individual disease progression. This synergistic effect happens particularly in the presence of hypertension and dyslipidemia.

What this adds to what is known:

- Multimobidity is not only about pairs of diseases, but also about how multiple diseases in patients interact and how this interaction changes over time. For the first time this proper perspective on multimorbidity is described and analysed using the new technique of multi-level temporal Bayesian networks. This new method not only supports finding multiple associations and how these change over time, but also which of these represent a direct association or a confounder, and how factors indirectly influence each other.

Implications:

- Whereas standard multilevel regression methods are very useful to explain a single disease with respect to a set of patient and practice related observable variables, multilevel Bayesian networks allow exploring the joint distribution of multiple diseases and their interactions, which is highly relevant in multimorbidity research.
- Clinical guidelines for patients with multimorbidity can be improved when the advice incorporates all the patient's specific characteristics. Since the network in the methodology used here can be personalised for a specific patient, it provides a valuable tool for the development of such tailored clinical guidelines.

Although systematic reviews [26, 39, 125] have given insight into the rates of cross-sectional co-occurrence, the progression over time of interactions between chronic cardiovascular diseases and related disorders is sparsely documented [61]. More insight into such interactions would help in personalising the therapeutic management of patients with multimorbidity. As clinical knowledge is mostly organised around single diseases, that knowledge may not be fully applicable to patients with multimorbidity [17, 11]. The care of patients with multimorbidity can be improved by any method that tailors the advice to the patient's specific characteristics [69].

Recent explorations of patient data from primary care registries to quantify associations between chronic disorders, have shown these to be valuable for obtaining a broad picture of multimorbidity [11, 77, 225]. In this chapter we explore such registries to assess three aspects of the joint progression of chronic cardiovascular multimorbidity: 1) its dependency on the practice's urbanity, 2) the synergistic effects between disorders when they evolve over time, and 3) the progression of the overall multimorbidity rate.

Patient data in primary care registries are often clustered by practices, introducing particular biases in the patient's diagnosis due to practice related effects. For example, the urbanity of the practice's area or the physician's experience. Multilevel regression analysis is the standard method of choice in these situations [83]. However, it does

not allow analysing multiple disease outcomes simultaneously. Therefore, we adopted the method of *multilevel Bayesian networks* (MBNs) that does offer such support; see Chapter 6 for details. When used for the analysis of temporal data, the advantage of an MBN is that the disorders and their interaction are all treated as uncertain. The representation goes beyond showing how pairs of disorders are associated to each other. Furthermore, we can extend an MBN to analyse multiple outcomes at multiple time-points. The latter gives rise to *multilevel temporal Bayesian networks*, or MTBN for short.

In summary, we developed a multimorbidity model that yields a much better insight into interactions, progression over time, and the accumulation of chronic disorders than existing statistical models are able to provide. Moreover, we show that posterior probabilities computed from the model at follow-ups can be tailored to any set of conditions present at baseline, which can provide valuable input for *personalised* clinical decision-making.

7.2 METHODS

7.2.1 *Data collection*

The data used for analysis were obtained from the Netherlands Information Network of General Practice (LINH); see Chapter 3 and Appendix A for further detail on the data and ICPC codes of the disorders used in the analysis. Here, longitudinal data of approximately one and half million patient years, covering the decade 2002-2011, from patients aged over 35 years, were used in our analysis. Patient data is available for the whole time frame, unless patient moved out of the practice or the practice itself opted out. Lab results and medication were not always consistent with the diagnoses present in the LINH database. They were corrected as described in Chapter 3.

We used the definition of a 'chronic disorder' given by O'Halloran [144], which in turn was based on the international classification of primary care (ICPC) codes. Principally, our focus is on chronic cardiovascular diseases and related disorders, and in our model we included the following chronic disorders: obesity, hypertension, lipid disorder, diabetes mellitus, heart failure, stroke, ischaemic heart disease, retinopathy, and nephropathy. The first three disorders are seen as health risks. Previous research has indicated that also some *non*-cardiovascular comorbidity is associated with cardiovascular disorders [11, 211]. Therefore, the diagnoses of other chronic non-cardiovascular disorders were modelled as well, but only as a single variable.

7.2.2 *Statistical analyses*

As the patient data in the LINH dataset were obtained from several general practices, differences among those general practices may have a confounding effect on the probability distributions. Taking into account the hierarchical structure into statistical models demands for a multilevel approach. We used MBNs in our analysis [105]. Bayesian networks provide a powerful framework for the representation of knowledge and reasoning under uncertainty [149], and they have had a significant impact on the modelling

and the analysis of medical data [120]. The statistical relationships in such models can be learned from patient data. In Chapter 6, it was shown that when applying MBNs to a set of hierarchically structured disease variables, the outcome of multiple diseases can be very well predicted using a MBN. With a receiver operating characteristic (ROC) curve it was demonstrated that a single MBN outperformed the use of multilevel regression models for each disease separately.

The use of a network-based approach to human disease, so-called *network medicine*, was recently acknowledged to be useful in researching complex disease pathways [10]. In an MBN, the disease variables are also represented as vertices in a network, but the associations have a direction and probabilistic associations are represented by directed edges. When there is an edge from vertex i to vertex j, then i is called a *parent* of j, and j is called a *child* of i. Although we cannot assume that these edges represent true causality, often they do. Temporal edges always point from the past to the future, and here a causal interpretation is even more natural. Each vertex is associated with a multinomial probability distribution for each configuration of the parent vertex. The interactions or moderating effects between parent vertices on their common child are therefore captured in these local probability distributions.

In the MBN used for this chapter, we modelled the patient's status in terms of the predefined disorders at baseline using the first five years of the data in retrospect, with a registration minimum of three years. The population used here is a fixed cohort of patients that were alive at baseline. The disease status three and five years after the baseline was also included in the model, although patients might have been deceased or moved to another practice at that time. Building the complete MBN required two major steps:

1. Specifying the qualitative nature of the relationships in the network.
2. Specifying the local probability distributions of the disease variables.

For step 1 all the disorders were represented as binary variables, i.e., the disorder is present yes or no. Age was discretised into four age groups. The subnetworks constituting the resulted three time slices were then connected in such a way that each disorder variable had at least one directed edge to the same variable in the next time slice. By doing this, the MBN is transformed into an MTBN, i.e., a multilevel, *temporal* Bayesian network. The urbanity of the practice was operationalised to control for several potential confounding factors by including it as a higher level variable in the MTBN.

Although the relationships between the disease variables, expressed by the edges in the network, can be specified by the user, we instead learned these relationships from the data. This allows for revealing yet unknown relationships. Both the relationships between disease variables within a time slice and between different time slices were learned using a score-based searching method in the statistical R package *bnlearn* [185]. We assumed that the qualitative nature of associations between disease variables do not vary over time, i.e., the network structure remains the same for each time slice. To ensure a multilevel structure and specific medical knowledge, dependencies between disease variables can be secured or avoided through black- and whitelisting. For example, if an association between a demographic variable C and a disease variable D was found by the structure learning algorithm, the corresponding edge in

the network clearly should be $C \rightarrow D$. Therefore, all possible edges $D^i \rightarrow C^j$ were blacklisted.

For step 2 we assumed that, although the network structure remains the same for each time slice, the parameters of the local probability distributions were allowed to change over time. The latter condition is known as the condition of non-stationarity. Once the complete structure was determined, the local probability distributions were estimated using one thousand bootstrapped samples from the dataset, i.e., we computed $P(D_t^i \mid \mathtt{parents}(D_t^i))$ with $D_t^i = 1$ if disease i is present at time t and $D_t^i = 0$ otherwise. Note that $\mathtt{parents}(D_t^i)$ is determined by the network structure; it can contain disease variables of several types, e.g., age, a health risk factor, or another disease, from both the current time slice and the previous time slice. The variance induced by the urbanization level (a practice level variable), age, and gender, in the multilevel model was explained by using Markov Chain Monte Carlo simulation in WinBUGS [195]. The learned structure and parameters were loaded into the software package SamIam (© UCLA) for probabilistic inference.

The total number of disorders present simultaneously in patients, i.e., the multimorbidity rate, was calculated using variables M_t that kept track of the number of disorders present in time slice t. The probability distribution of these variables is deterministic, i.e.:

$$P(M_t = q \mid D_t^1, D_t^2, \cdots, D_t^n) = \begin{cases} 1 & \text{if } \sum_{i=1}^{n} D_t^i = q \\ 0 & \text{otherwise} \end{cases} \tag{4}$$

The value q then represents the number of simultaneously present disorders. See Appendix A for more detail on the implementation of MTBNs.

Once the local probability distributions were determined we were able to answer the questions mentioned in the introduction. The urbanity effects were derived by conditioning on a specific value of the corresponding vertex in the network. Secondly, for each time slice we determined whether cumulative incidence rates of disease combinations significantly deviated by increased occurrence from what might be expected from individual cumulative incidence rates, assuming statistical independence. Mathematically this can be expressed for two disorders i and j as:

$$P(D_t^i, D_t^j \mid R_t) \gg P(D_t^i \mid R_t)\ P(D_t^j \mid R_t) \tag{5}$$

with R_t the a set of health risks R_t^k present at time t. Since the size of the studied patient population favours reaching significance easily, we also examined the clinical importance of such deviations.

By conditioning on the multimorbidity rate M_t in a specific time slice the model allowed us to predict the multimorbidity rate in the next time slice, which is mathematically expressed by:

$$P(M_t = q \mid M_{t-1} = r) \tag{6}$$

We are particularly interested in the probabilities for $q > r$, because if these are large, the number of simultaneous disorders increases with time. These probabilities can be biased due to possible disease shifts, i.e., the patient acquires a new disorder but also

loses one, keeping the multimorbidity rate equal. To evaluate this effect we calculated how much acquired disorders sustained in the next time slice, mathematically:

$$P(D_t^i = 1 \mid D_{t-1}^i = 1) \tag{7}$$

If these probabilities are close to one, the effects of disease shifts are considered to be minimal.

7.3 RESULTS

The final MTBN consists of three time slices modelling chronic cardiovascular disease progression. The associations in the MTBN are summarised in Table 7.1. The complete network structure representing all the parent-child relations, and thus the qualitative nature of the underlying multivariate distributions, is available in Appendix A, together with the pseudo-code in WinBUGS for parameter estimation. Evidently, age had a significant association with all other variables. However, gender did not had a significant association with CVD, except for ischaemic heart disease.

Data of a total of 182,396 patients were used for analysis. The median and mean age of the patients at baseline were 53 and 55 years, respectively. At the end of the five year follow-up, 8.5% of these patients had dropped out of the registry. This happened because of death or patients moved to a nursery home or practice not present in the registry. Their disease status until this event was included in our analysis.

For all health risks and chronic disorders that were incorporated into the model, cumulative incidence rates, with their standard errors, were estimated for each time slice. These were differentiated for age and urbanity, and to examine the model's validity further, we also made model predictions for diabetics and non-diabetics. Besides the individual rates, we also calculated rates of comorbidity patterns. Details on demographics and probability estimations are available on-line. The effect of urbanity on disease probabilities, corrected for age and gender is shown in Figure 7.1.

Table 7.2 shows the evaluation of synergies, as defined in Equation (5), in which cumulative incidence rates of the comorbidity patterns are compared with the rates of single disorders. As dyslipidemia and hypertension are the major predictors of cardiovascular morbidities, we computed the conditional probabilities in the absence and presence of these conditions from the MTBN. Some of the comorbidity patterns deviate significantly from the expected values. For example, at five-year follow-up the probability of ischaemic heart disease and heart failure together is 5.4% when both dyslipidemia and hypertension are present. However, using Equation (5), the product of their individual rates is only 2.9%. The true incidence is thus almost twice as high, which indicates an interaction between the two disorders in relation to hypertension and dyslipidemia. This phenomenon can be found for several comorbidity patterns through each time slice. Since probabilities of disease combinations are relatively low, we consider absolute increments, rather than relative increments, where an increment of 0.5% is considered to be of clinical relevance.

Figure 7.2 shows the multimorbidity rates, as defined in Equation (6). For patients having one or more health risks, the probability of obtaining a new health risk is relatively low, in comparison to the presence of other cardiovascular disorders. In that case, the cumulative incidence rate rapidly increases with the number of conditions present

Chronic Disease	Associations known from the literature	Associations learned from the data	
		Direct*	Indirect**
diabetes mellitus	age, dyslipidemia, hypertension, ischaemic heart disease, heart failure, nephropathy, retinopathy obesity, stroke [176]†, practice [140]	age, dyslipidemia, hypertension, ischaemic heart disease, heart failure, nephropathy, retinopathy,	practice, obesity, stroke
ischaemic heart disease	age, gender, obesity, dyslipidemia, hypertension, diabetes mellitus, heart failure, stroke [156]†, retinopathy [173], practice [77]	age, gender, dyslipidemia, hypertension, diabetes mellitus, heart failure	practice, obesity, stroke, *nephropathy*, retinopathy
heart failure	age, obesity, dyslipidemia, hypertension, diabetes mellitus, ischaemic heart disease, nephropathy, stroke [129]†	age, hypertension, diabetes mellitus, ischaemic heart disease, stroke, nephropathy	*practice*, obesity, dyslipidemia, *retinopathy*
stroke	age, obesity, dyslipidemia, hypertension, diabetes mellitus, ischaemic heart disease, heart failure [156]†, practice [77]	age, dyslipidemia, hypertension, heart failure	practice, obesity, diabetes mellitus, ischaemic heart disease, *nephropathy*, *retinopathy*
nephropathy	age, gender, hypertension, diabetes mellitus, ischaemic heart disease, heart failure, stroke [220], retinopathy [122]	age, hypertension, diabetes mellitus, heart failure	*practice*, *obesity*, dyslipidemia, ischaemic heart disease, retinopathy
retinopathy	age, dyslipidemia, hypertension, diabetes mellitus [176]†, ischaemic heart disease [173], nephropathy [122]	age, hypertension, diabetes mellitus	*practice*, *obesity*, dyslipidemia, ischaemic heart disease, *heart failure*, nephropathy

Table 7.1: Associations between cardiovascular diseases, known from the literature (clinical guidelines are marked with a †), and learned from the data. * Direct associations correspond with an edge between diseases in the MTBN. ** Indirect associations correspond with diseases that have another disease between them or share a common child or parent in the MTBN. Italic written associations were not directly clear from the used literature.

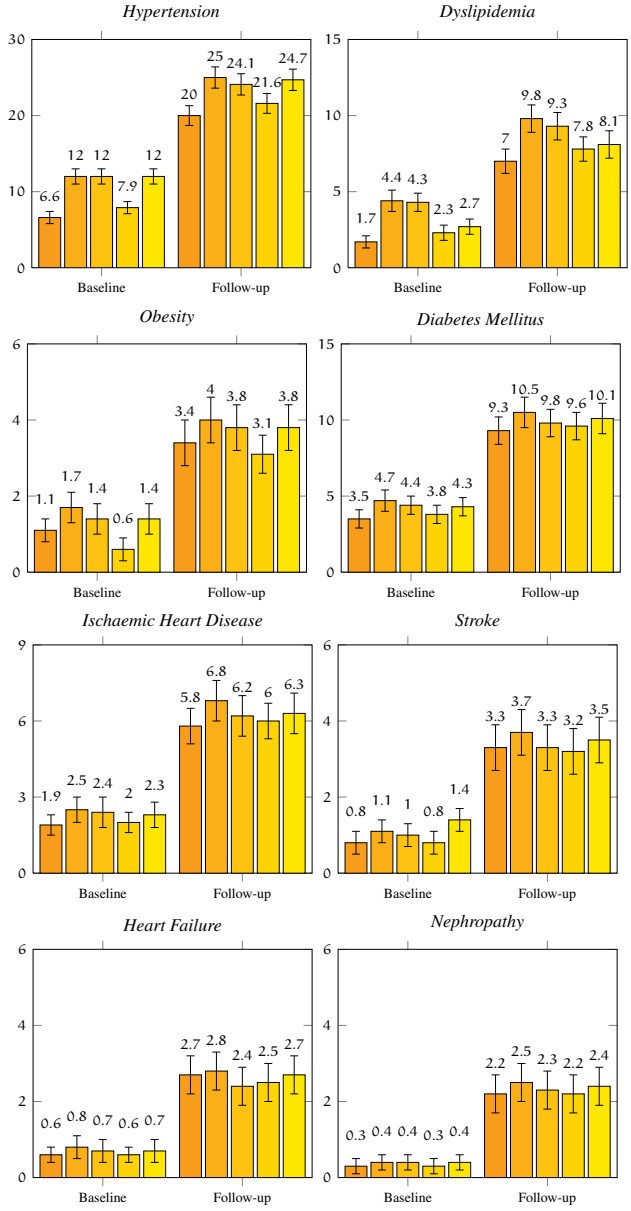


Figure 7.1: The effect of the urbanization level, varying from very high urban areas ($>$ 2500 addresses per km^2, orange) to rural areas ($<$ 500 addresses per km^2, yellow), on cumulative disease incidence at base-line and five year follow-up. Numbers are corrected for age and gender and provided with a 95% confidence interval.

	BaseLine				3 years follow-up				5 years follow-up			
Risk Factors	None	DL	HT	DL+HT	None	DL	HT	DL+HT	None	DL	HT	DL+HT
Comorbidity												
DM+IHD	0.2	2.6	1.8	6.2	0.7	5.8	4.4	11.2	1.0	6.8	5.4	14.0
DM+HF	<.1	0.5	0.7	1.3	0.4	1.6	2.1	3.9	0.5	2.5	3.0	5.0
DM+NP	<.1	0.3	0.4	0.9	0.3	1.1	2.0	3.8	0.5	1.7	3.1	5.0
DM+ST	<.1	0.7	0.7	2.6	0.2	1.9	2.1	5.3	0.4	2.4	2.9	6.4
DM+RP	<.1	0.1	0.1	0.2	0.1	0.2	0.2	0.3	0.1	0.3	0.3	0.4
IHD+ST	<.1	0.6	0.4	1.7	0.2	1.6	1.3	3.8	0.3	2.3	2.0	4.9
IHD+NP	<.1	<.1	0.2	0.5	0.1	0.6	1.0	2.0	0.2	1.1	1.7	3.4
IHD+HF	<.1	0.8	0.9	1.8	0.4	2.1	2.2	3.9	0.6	2.8	3.2	5.4
ST+HF	<.1	0.1	0.3	0.4	0.2	0.5	0.9	1.4	0.3	0.9	1.4	2.2
NP+HF	<.1	<.1	0.1	0.2	0.1	0.4	0.9	1.1	0.3	0.6	1.5	2.0

Table 7.2: Estimated probabilities of having comorbid combinations of chronic cardiovascular diseases at baseline, and after three and five years follow-up, under condition of the presence or absence of certain health risks. Abbreviations: HT=hypertension; DL=dyslipidemia; DM=diabetes mellitus; IHD=ischaemic heart disease; HF=heart failure; ST=stroke; NP=nephropathy; RP=retinopathy. Results are shown in percentages. The yellow part of the circle represents the expected value based on individual rates, the surplus is coloured in orange or red (see also Equation (5) of this chapter). Red circles represent cumulative incidence rates which deviate significantly ($p < 0.001$) from the expected values and have a clinical importance as well (absolute increase $> 0.5\%$). They indicate the clinical significant interactions.

Chronic Disorder	$t_0 \rightarrow t_1$	$t_1 \rightarrow t_2$
Obesity	40	33
Dyslipidemia	78	90
Hypertension	95	97
Diabetes Mellitus	95	91
Ischaemic Heart Disease	94	98
Heart Failure	95	99
Stroke	90	91
Nephropathy	98	99
Retinopathy	99	99
Other	66	90

Table 7.3: Persistence of individual chronic diseases, i.e., the probability (in percentages) of a disorder being present at follow-ups under the condition that this disorder was present in the previous time slice (see also Equation (7) of this chapter). t_0=baseline, t_1=follow-up 3 years, and t_2=follow-up 5 years.

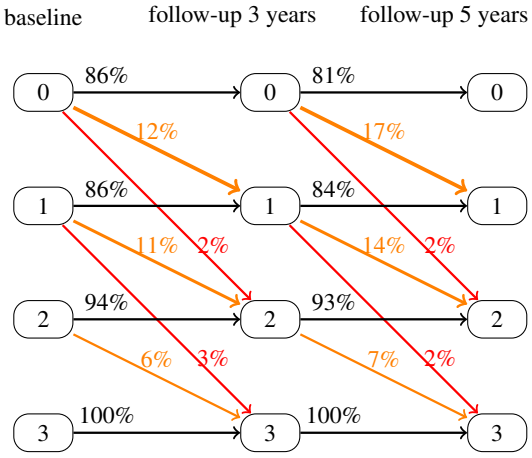
in the previous time slice. For example, when having two disorders at baseline, the probability of obtaining one or more cardiovascular disorders after three year follow-up is approximately 19%+3%=22%, following the edges from vertex two (at baseline) to vertex three and four plus (at three years follow-up) at the right-hand side of Figure 7.2. From the remaining 78%, those who attracted no new cardiovascular disorder within three years, another 32% gets one or more disorders at five years follow-up, making the total probability 47%.

Table 7.3 shows the persistence probabilities of individual disorders at follow-ups, as defined in Equation (7). For the majority of the disorders over 90% sustained in the next time slice. Obesity is the major exception on this.

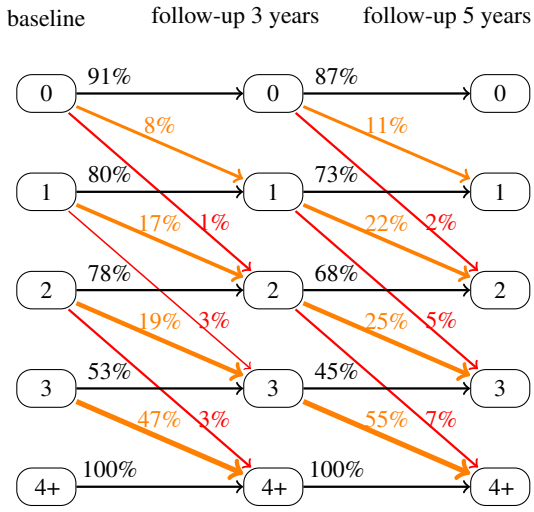
7.4 DISCUSSION

In this study, the new method of multilevel temporal Bayesian networks was used to precisely capture the *qualitative* and *quantitative* time course of chronic cardiovascular multimorbidity in general practices. Bayesian network methods have not been used before in multimorbidity analysis. Although the discovered network dependences are sometimes similar to the associations described in medical literature, discovered by standard statistical means, the global picture of how chronic disorders and risk factors influence each other, as represented by a multilevel Bayesian network, is new.

It gives an overview of *direct* and *indirect* associations and it quantifies transition rates, all in one representation. The results obtained are discussed in more detail below.



a) # health risk determinants



b) # chronic cardiovascular diseases

Figure 7.2: Transition probabilities of a) health risks, and b) chronic cardiovascular diseases (see also Equation (6) of this chapter). Health risks are obesity, dyslipidemia, and hypertension. Chronic cardiovascular diseases are diabetes mellitus, ischaemic heart disease, heart failure, stroke, retinopathy, and nephropathy. The left (black) percentages and lines represent patients who do not acquire a new determinant or chronic disease within the next time slice, as where the middle (orange) and right (red) percentages and lines represent patients acquiring respectively one or two new health risks or chronic diseases within the next time slice.

7.4.1 *Evaluation of the network structure*

The associations in summarised in Table 7.1 are compared with current knowledge reflected in recent clinical guidelines and the medical literature. A distinction is made between direct and indirect dependences, something not possible when using standard statistical methods. For example, represented in the MTBN there is an indirect association between dyslipidemia and heart failure via, i.e., conditional on, ischaemic heart disease and hypertension. This is in line with clinical guidelines on heart failure, which states that there is no reason to prescribe a statin in the absence of an ischaemic cause of heart failure [156].

Other disorders share a common parent in the network structure, e.g., both retinopathy and nephropathy share diabetes mellitus and hypertension as a common parent. The investigation of either of them in the presence of the other, is thus only of beneficial value if either diabetes mellitus or hypertension is also present.

The analysis shows that gender-induced associations are insignificant or small, and in case it is small, it is of little clinical significance. The exception on this is ischaemic heart disease, where it is well established that gender is significant [203].

The comorbid associations between cardiovascular related chronic disorders and other chronic disorders [211], are also recognised in our model; obesity, hypertension, dyslipidemia, and diabetes mellitus are associated with *non*-cardiovascular disorders. Moreover, the temporal associations show that diabetes mellitus is a direct predictor of such disorders in follow-ups.

Although the network structure indicates that obesity is a good predictor of hypertension and dyslipidemia, we observe that direct associations between obesity and conditions other than hypertension and dyslipidemia are missing. For example, one would have expected an edge from obesity to diabetes mellitus, because it is well known that an elevated body mass index or waist circumferences is associated with diabetes mellitus [176]. Hence, the effect of obesity on other conditions is probably underestimated in the data explored, in particular due to low persistence at follow-ups in the database. The latter does not mean that most patients actually lose weight, but that registries do not properly keep track on this matter. However, if we leave obesity out of the model, it has little effect on the structure and the associated probabilities are minimally affected (data not shown).

7.4.2 *Quantitative analysis*

The prevalences of morbidity and comorbidity patterns at three-year follow-up are comparable to prevalences obtained from previous studies using LINH data, [215] and earlier results within the Netherlands [207]. The associations between age, diabetes and cardiovascular multimorbidity are quantitatively well recognised by the model. The multilevel approach allowed us to differentiate probabilities for practice related variables. Where other researchers showed an association between multimorbidity and socio-economic status [177, 11], we modelled the urbanity of the practice along with the disease variables. In case of obesity, hypertension, dyslipidemia, diabetes mellitus, and ischaemic heart disease, the urbanization level had a significant effect on the prevalence. In these cases, the cumulative incidences of moderate and high urban areas

were mostly above average, whereas these incidences of low and very high urban areas were mostly below average. Rural areas show on average incidences comparable to the overall average.

The effect of multiple cardiovascular risk factors was already outlined in [87], indicating that the five-year cardiovascular risk can go up to 44% when certain risk factors, e.g., hypertension, total cholesterol, smoking, high density lipoprotein, gender, diabetes, and high age, are present. Their results indicate, for example, that a smoking male patient over 60 years with diabetes, high total cholesterol and low high density lipoprotein, has a 5-year cardiovascular risk of 22% and 44%, for low and high systolic blood pressure respectively. We can do more or less the same exercise as described above: the 5-year cardiovascular risk for a male diabetic living in a rural area, aged between 65 and 80 years, and having a lipid disorder, is 28% for non-hypertensive patients and 55% for hypertensive patients.

However, the multiple-risk attributions in their approach could only be derived by adding the risk factors consecutively in a specific order. In our model there is no restriction on the number of disease variables used as predictor and the number of diseases variables being predicted. For example, the 5-year cardiovascular risk of *two or more* new diseases for is 10% and 23% for the patient described above. It also implies that we can condition on a specific cardiovascular disease already present. When conditioning on heart failure the effect on cardiovascular risk is the highest, e.g., a male hypertensive diabetic with dyslipidemia, living in a rural area, aged between 65 and 80 years, with heart failure already present at baseline, has a 5-year risk of 71% to obtain another cardiovascular disease, and 27% for two or more diseases.

Although any other cardiovascular risk score also represents a personalisation, the major difference with an MTBN is that in an MTBN not all disease variables need to be known. In fact an MTBN captures all predictions for *any* disease variable within the model for *any subset* of the remaining variables. This means that one can also reason the other way around: given the presence of certain diseases, one can make predictions about the presence of specific risk factors, e.g., hypertension or lipid disorders are more likely to be present in the presence of cardiovascular diseases. In Figure 7.3 an example is shown of a personalisation, indicating that multimorbidity at baseline predicts future multimorbidity better than the demographics do. This is in line with the idea that the patient's *biological* age is of more importance in relationship to morbidity than *chronological* age [133], and the relation between frailty and the accumulation of deficits [172].

In summary, an MTBN can be used to make predictions for multiple diseases in many ways. To our knowledge this is new in multimorbidity research, and Table 7.2 only reflects a particular personalisation of cardiovascular risk. It reveals multiple interactions between chronic cardiovascular diseases and related disorders, which occur more frequently at follow-ups in comparison to baseline. We shall not discuss every interaction in detail, but it appears that the presence of hypertension or dyslipidemia are necessary preconditions for finding clinical significant interactions. For example, the combination of ischaemic heart disease and heart failure is much higher than expected after three- and five-year follow-up. Alternatively, although cumulative incidences increase over time, they behave as expected, e.g., in the case of ischaemic heart disease in

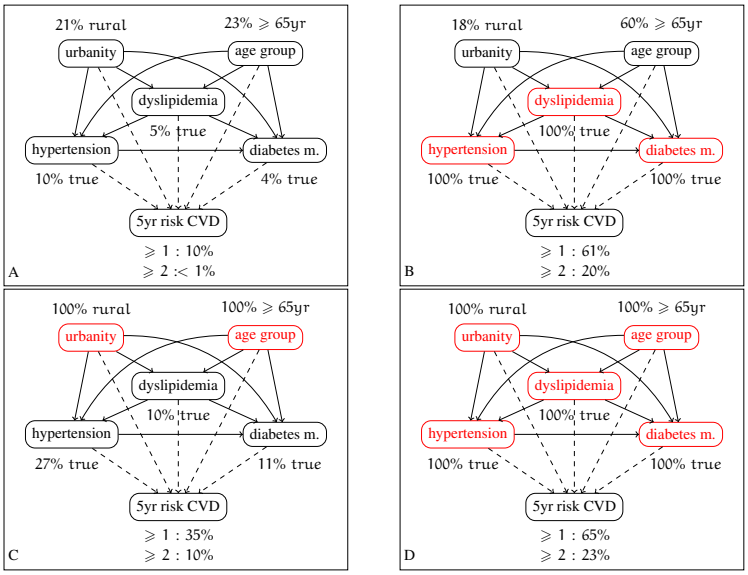


Figure 7.3: personalisation of a subset of the disease model. Straight lines represent the atemporal associations at baseline, and dashed lines represent the temporal associations between baseline conditions and the five year cardiovascular risk. Subfigure (A) shows the 5-year risk on a cardiovascular (related) disease (ischaemic heart disease, heart failure, stroke, retinopathy, or nephropathy) of the general Dutch population aged over 35 years. This probability is 10%, and within this population we have the following prior probabilities of risk factors: 21% lives in a rural area, 23% is older than 65 years, 10% has hypertension, 5% has dyslipidemia, and 4% has diabetes mellitus. The subfigures (B), (C) and (D) show how probabilities change when specific information of the patient is incorporated. (B) represents patients with hypertension, dyslipidemia and diabetes at baseline, but without knowledge about the patient's demographics. The 5-year risk has increased to 61%, and it is estimated that 18% lives in a rural area and 60% is over 65 years. (C) represents patients older than 65 years, living in a rural area, but without knowledge about the patient's disease status. The 5-year risk on CVD has increased to 45% and it is estimated that cumulative incidences of risk factors are doubled or more. (D) is the combination of (B) and (C), showing that the 5-year risk has gained only a little with respect to (B).

combination with stroke. This fits with the fact that only an indirect association exists in the model.

Another new aspect of our model are the temporal associations. They cause the incidence of chronic disorders to rise quickly over time. In particular, the probability of acquiring at least one new chronic cardiovascular disorder increases with the number of chronic cardiovascular disorders already present, regardless of age (Figure 7.2). At three years follow-up this is respectively 9%, 20%, 22% and 47%, for zero to three disorders present at baseline. At five years follow-up this has increased to respectively 21%, 42%, 47% and 76%. In reality, these numbers can be even higher due to disease shifts. Since the probability of sustaining a chronic disorder is at least 90%, for the ones we used in our model, we believe this effect is minimal.

Cross-sectional research of other registries show that the prevalence of multimorbidity can be up to 90% [207, 224, 53, 11]. Moreover, 80% of the elderly patients with heart failure face at least *four* chronic comorbidities [211]. These numbers are comparable with the prevalences of cardiovascular comorbidity at follow-ups retrieved in our model. However, the prevalence of cardiovascular multimorbidity, in particular for diabetes mellitus, ischaemic heart disease and stroke, is much lower at baseline in our model. This indicates the importance of the temporal dimension; estimates cannot be directly extrapolated to follow-ups, e.g., by using the prevalences of a higher age group.

7.4.3 *Strength, limitations and implications*

The major strength of the results is that the obtained MTBN allows analysing several aspects of multimorbidity in a single model. Our research encompassed an analysis of data obtained from public health registries, and because of the size of the data set used, significant results could be established. Although the data used here contained more patients and more disease variables as mostly present in controlled studies, it also contains more noise and typically involved more preprocessing. Controlled studies are relatively small in size and often exclude patients with multimorbidity. There are some exceptions on this, e.g., recently a cohort study of nearly 15,000 elderly people had a focus on the epidemiology of chronic diseases using a variety of biomarkers and non-invasive measurements [80]. But the majority of its research has a single-disease focus. We recommend to apply Bayesian networks here as well to discover the coherence between the multiple biomarkers and disorders present in such studies.

Several aspects of the results could have been analysed by alternative methods. For instance, multi-state models could have been used to analyse the transition rates in the multimorbidity number, a separate multilevel regression model for each disease to investigate the urbanity effects, and a chi-squared test to see if the joint prevalence of two conditions is higher than would be expected. However, investigating more complex interactions, like in Table 7.2, would require logistic regression with added *interaction terms* [86]. Logistic regression demands building a separate model for each disorder; in this way, the insight into the qualitative nature of the interactions between the disorders would be lost.

With an MTBN we also avoid the redundancy that is obtained when using multiple separate regression models for each disease. For example, if we regress disease D on

disease D' in one model and in another model D' on D, we obtain two parameters for the same association. This could produce certain ambiguity because the two models do not necessarily have to provide the same odds ratio for that pair. In summary, the MTBN used here allows analysing all the results presented in this chapter without losing any of the epidemiological coherence between all the disease variables. To our knowledge, this is new in multimorbidity research, and there is no *single* alternative that analyses multimorbidity the way we did.

There are some aspects of registries that introduce a certain bias. Patients that did not visit their physician within the used time frame are not included in the data. Although we explored a decade of patient data within a time frame of ten years and patients of age 35 and above, making the proportion of missing patients likely to be very low, prevalences are probably slightly overestimated. On the other hand, in a public health registries data are missing and there are also incorrectly coded diagnoses, implying that prevalences might also be underestimated. Clinical guidelines often recommend specific additional investigations making it that certain disorders are discovered more likely than others. For example, retinopathy in a diabetic may only be discovered because of the recommendation mentioned in the guideline of visiting an ophthalmologist.

In our results, there is a considerable prevalence change between the baseline and follow-ups. Partly this is due to the fact that the population has aged five year. On the other hand, the absence of a diagnosis is interpreted as the absence of the corresponding disease, however, certain pathophysiology could already be present at baseline without knowing it. This delay in diagnosis also makes longitudinal associations between disorders less detectable.

7.5 CONCLUSIONS

Several attempts have been made in the literature to capture prevalences of multimorbidity. Lately, electronic databases of general practices are used more and more to quantify these numbers on a larger scale, but there is no clear method that fully describes how multiple disease evolve over time. Traditional statistical techniques are very useful to evaluate a single-disease framework by which most medical care, research, and education is configured. However, a multiple-disease orientation requires a more complementary strategy.

In that respect, the MTBN used here is a valuable step forwards in multimorbidity research. It combines the advantages of a temporal Bayesian network together with a multilevel analysis, and it was able to discover complex multimorbidity patterns of chronic diseases within healthcare data. First, the model was shown to be valid by comparing known disease interactions for diabetes mellitus with those present in the network structure. Second, several new disease interactions, qualitative and quantitative, for three and five years follow-up were discovered, showing that cumulative incidence rates are accelerated in the presence of multimorbidity. Especially, the presence of conditions such as hypertension, dyslipidemia, and diabetes mellitus accelerates cardiovascular risk significantly.

Here we only discussed the most significant results that can be obtained from an analysis of the MTBN model. The model itself can be used to extract many other

relevant conclusions. We conclude that Bayesian network models make the analysis and visualization of the interactions between chronic disorders and their evolutionary course more comprehensive than traditional statistical techniques. They can be used to answer a variety of clinical and epidemiological questions without losing the context of these dependences out of sight. This is of great importance in the management of multimorbidity and the aim to adopt personalised clinical guidelines. The next step in multimorbidity research might be to address mortality rates, differentiated for cardiovascular and non-cardiovascular chronic diseases, in the same way as was done here for multimorbidity rates.

8

QUALITATIVE CHAIN GRAPHS AND THEIR APPLICATION

ABSTRACT

For many problem domains, such as medicine, chain graphs are more attractive than Bayesian networks as they support representing interactions between variables that have no natural direction. In particular, interactions between variables that result from certain feedback mechanisms can be represented by chain graphs. Using qualitative abstractions of probabilistic interactions is also of interest, as these allow focusing on patterns in the interactions rather than looking at the numerical detail. Such patterns are often known by experts and sufficient for making decisions. So far, qualitative abstractions of probabilistic interactions have only been developed for Bayesian networks in the form of qualitative probabilistic networks. In this chapter, such qualitative abstractions are developed for chain graphs with the practical aim of using qualitative knowledge as constraints on the hyperspace of probability distributions. The usefulness of qualitative chain graphs is explored for modelling and reasoning about the interactions between diseases.

8.1 INTRODUCTION

Probabilistic graphical models (PGMs) have been shown to be convenient and intuitive formalisms to capture the probabilistic independence information in many application fields. In a PGM, random variables are modelled as vertices connected by edges in a graph. These connections reflect the probabilistic dependences and independences between variables and one can associate a probability distribution to the graph that is faithful in some way to the dependences and independences. Popular PGMs include models based on undirected graphs (UGs), i.e., *Markov networks*, and based on directed acyclic graphs (DAGs), i.e., *Bayesian networks* [149]. However, both undirected and directed graphs have certain undesirable limitations when representing independence information for an actual problem domain. Hybrid graphs, containing both directed and undirected edges, such as *chain graphs*, offer an elegant generalisation of both Markov and Bayesian networks [108].

A chain graph (CG) uses potentials rather than straight probabilities to represent the probability distribution of variables and is, therefore, often seen as a blackbox model. Nevertheless, chain graphs have been shown to model equilibrium systems [107], which occur in many areas including biology, physics, chemistry, and economics. In fact, it

was shown that particular sets of conditional independence statements, which cannot be modelled by a Bayesian network, can indeed be modelled with a chain graph; the ideal gas law and the price and demand model in economics are examples [34].

On the other hand, Bayesian networks have the advantage that both structure and parameters can be assessed from either expert knowledge, data, or both, which renders Bayesian networks whitebox rather than blackbox models. For the more expressive chain graphs, it is much more difficult to exploit human knowledge in assessing their parameters, and, as a consequence, these models do not share all the advantages of Bayesian networks as whitebox models. One of the aims of the research described in this chapter is to come up with ways to move chain graphs closer to whitebox models, in particular by the use of qualitative probabilistic abstractions.

Probabilistic information is available in different forms, ranging from numerical, quantitative probabilistic values (possibly with a confidence interval) to qualitative information. Qualitative abstractions of Bayesian networks, called *qualitative probabilistic networks* (QPNs) [221], offer a useful method for exploiting qualitative constraints in assessing probabilistic information. Qualitative information in QPNs may consist of qualitative influences and synergies, and independence information. While it is well known that QPN theory has its limitation when it comes to qualitative reasoning – the main reason why QPN theory is not used in actual systems – qualitative knowledge may be quite useful when looked at as offering constraints that should be taken into account when estimating a probability distribution. Some algorithms have been proposed in the past to derive bounds [114] and qualitative influences [24] in the presence of both quantitative and qualitative knowledge, with applications in e.g. computer vision [35]. Furthermore, it has been proposed to derive marginal probability distributions in the presence of such hybrid knowledge [46]. If exact probabilistic information is not required, then such distributions, also called second-order distributions[1], provide insight into the domain and could, e.g., be used to make decisions.

In the next section, we will first argue why chain graphs provide a good starting point for modelling feedback mechanisms. Here we explore three realistic examples drawn from the medical field. The needed theoretical basis underlying the work presented in this chapter is provided in Section 8.3. In Section 8.4, we extend the known QPN theory towards chain graphs, which we call *qualitative chain graphs* (QCGs). In particular, we will formally discuss qualitative relationships, compare these to the relationships in QPNs, and prove their most important properties. In Section 8.5, we show that sign propagation, a qualitative variant of belief propagation, can be amended to qualitative chain graphs. In Section 8.6, we also demonstrate their usefulness in semi-qualitative reasoning and present experimental results supporting this claim. Although examples were drawn from the field of medicine, which offers a rich source of qualitative modelling, the results will be of value to many other domains. The work is rounded off by conclusions and plans for future research in Section 8.7.

1 In this context, a second-order distribution yields a distribution over all possible probability distributions that obey a set of qualitative constraints.

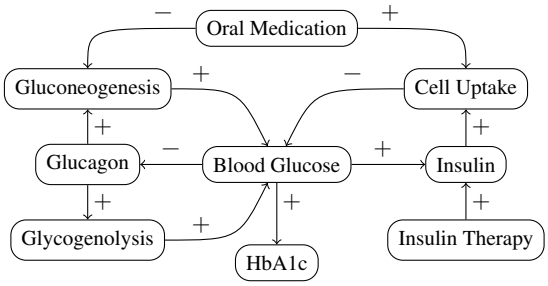


Figure 8.1: Single disease modelling – a graphical representation of physiological processes involved in blood glucose regulation.

8.2 MOTIVATION FROM THE MEDICAL FIELD

Many regulatory mechanisms within the human body, described by its physiology, can be seen as causal feedback systems, in which some kind of equilibrium setpoint – called *homoeostasis* – is maintained. Diseases can be conceived as a derangement of one or more regulatory mechanisms and treatments typically interact with these systems in non-trivial ways. In non-healthy people the equilibrium setpoint typically differs from the healthy people, but therapeutic interventions can reset the equilibrium setpoint to a state that is closer to that of the healthy people.

Example 1 concerns a simplified model of the blood glucose level regulation, showing how different agents, natural and pharmacological, have their role in maintaining the blood sugar homoeostasis. The representation here is often found in medical textbooks. A plus-sign typically represents stimulation of a process, and a minus-sign typically represents inhibition.

Example 1. *Blood sugar levels are regulated by negative feedback systems in order to keep the body in homoeostasis. High blood glucose levels stimulate the secretion of insulin by the pancreas, inducing glucose uptake in peripheral tissue. High blood glucose levels inhibit the secretion of glucagon by the pancreas, thereby also inhibiting glycogenolysis and gluconeogenesis, which both elevate blood sugar levels. Figure 8.1 shows a graphical representation of the blood glucose regulation, as typically used in medical textbooks. Glycated hemoglobin (HbA1c) is a marker of average blood glucose levels over the previous months, and thus provides a valid measurement of the blood glucose equilibrium. In diabetics, elevated glucose levels are caused by either an impaired insulin production (type I), or an insulin resistance of peripheral cells (type II). Possible solutions to re-establish a healthy equilibrium are insulin therapy or oral medication.*

The disturbance of the equilibrium of one physiological process might also alter the equilibrium setpoints of other regulatory systems, which might in turn induce new pathophysiology that deteriorates the patient's prognosis even further. In the interest of the physician, it is important to know the qualitative dynamics of such interactions, e.g., whether it is likely that a therapy for a specific disease gives rise to symptoms related to another (patho)physiological process. Example 2 shows the possible interactions

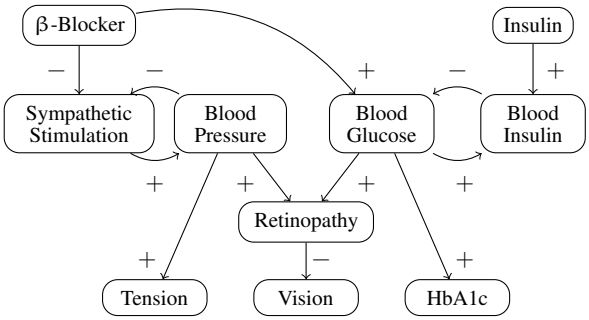


Figure 8.2: Multiple disease modelling – Interactions between multiple physiological regulation mechanisms: blood pressure regulation, blood glucose regulation, β-blocker therapy, insulin therapy; and some pathophysiological findings, e.g., hypertension, impaired vision, and a high blood glucose biomarker (Hb1Ac).

when two different therapies are administered: a blood glucose lowering therapy and an antihypertensive therapy.

Example 2. *A simplified abstract model of blood pressure and blood glucose regulation is shown in Figure 8.2. The blood pressure is regulated, among other physiologic processes, by a feedback loop with sympathetic stimulation. The same applies to blood sugar and insulin. Both high blood pressure and a high blood glucose level contribute to retinopathy, leading to impaired vision. Beta-blockers may stimulate the onset of diabetes mellitus [67].*

Another medical problem domain where associations without natural direction appear is the co-occurrence of two or more chronic diseases at the same time; this problem, described in detail in Chapter 3, is referred to as the problem of *multimorbidity*. Several attempts give a probabilistic classification of the different associations between diseases that occur when multiple diseases are present within one patient. An overview can be found in Chapter 4, which discusses multimorbidity in terms of association and direct causation.

Often, the associations between diseases occur because the diseases share the same pathophysiology, or the physiology is somehow related. When modelling such processes, Bayesian networks cannot adequately model the feedback mechanisms, and more expressive models are required. It is natural to model these associations by mixtures of undirected edges (lines) and directed edges (arcs), which is for example done in [204], even though the models presented in that paper are not given a (formal) probabilistic meaning. Example 3 shows a simplified feedback mechanism between diabetes mellitus and a lipid disorder. The variables of this model can be easily measured by a physician in general practice. Since we have patient data available coming from several general practices, this last example will be used as a running example throughout this chapter. This patient data is then used as input for the experimental results in Section 8.6.

Example 3. *Figure 8.3 shows an abstraction of the interaction between two diseases, i.e., diabetes mellitus and a lipid disorder, along with its typical blood measurements,*

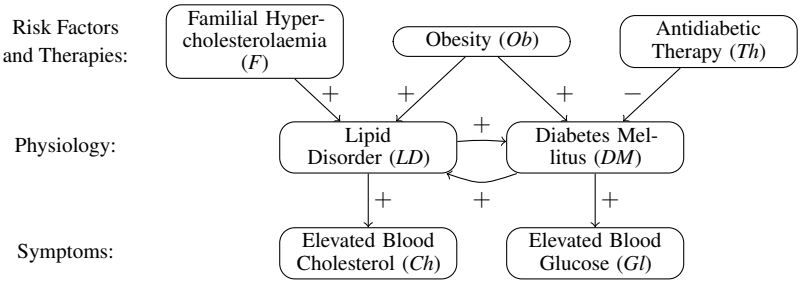


Figure 8.3: Running example – Schematic representation of the interaction between diabetes mellitus and lipid disorders, showing that between the diseases feedback exists within their pathophysiology.

two risk factors, i.e., obesity and familial hypercholesterolaemia, and a possible therapy for diabetes. It is assumed that there is feedback between the pathophysiology of both diseases, which is almost always in some kind of equilibrium. The relation between these pathophysiologies can be determined by the fact that the symptoms of each pathophysiology are linked to each other, i.e., elevated glucose levels are associated with elevated cholesterol levels, and given the current status of obesity and diabetes mellitus, the status of a lipid disorder is independent of the therapy for diabetes.

8.3 PRELIMINARIES

In this section, we introduce the necessary technical preliminaries used in the remainder of this chapter. In particular, we introduce chain graphs and their properties, in particular factorisation criteria and probabilistic independence. After this, we briefly introduce QPNs and their properties for qualitative reasoning.

8.3.1 *Notation*

In the following, we will denote random variables by, sometimes indexed, capital letters A, B and C and we write V, W, X, Y and Z for sets of random variables. For clarity of exposition, we will assume that each random variable A is a binary variable, which can take the values a ($A = \mathit{true}$) and $\bar{a}$ ($A = \mathit{false}$). Further, for notational convenience, we will sometimes write the singleton set $\{A\}$ as A, and, if X and Y are sets of random variables, then we will write XY instead of $X \cup Y$. Also, we write $X - Y$ for $X \setminus Y$. For example $X - AB$ is an abbreviation of $X \setminus \{A, B\}$. In probabilistic expressions, we adhere to the standard convention that $\{A_1, \ldots, A_n\}$ should be understood as the conjunction $A_1 \wedge \cdots \wedge A_n$, which we also sometimes use. A conditional probability distribution $P(X \mid Y)$ is defined as $P(X, Y)/P(Y)$, for positive $P(Y)$, and X is marginally independent of Y if $P(X \mid Y) = P(X)$. Finally, if we write an unbounded set of random variables X in a probability expression, then the expression is implicitly universally quantified over all configurations of X, for example, $P(a \mid X) = P(a)$ expresses that for all X (either x or $\bar{x}$ if it is a single variable) the probability of a given X is the probability of a alone.

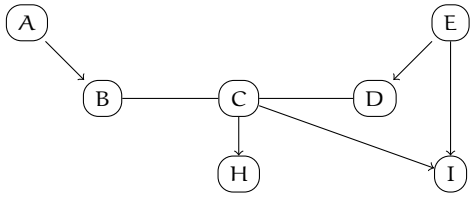


Figure 8.4: A chain graph G' used to illustrate some of the graphical concepts.

8.3.2 *Chain graphs*

INTRODUCTION A chain graph (CG) is a probabilistic graphical model that consists of labelled vertices, representing random variables, connected by directed and undirected edges. These models were originally introduced by Lauritzen and Wermuth [108]. Frydenberg [57] proposed a Markov property for these models, which is now known in the literature as the Lauritzen-Wermuth-Frydenberg (LWF) interpretation of chain graphs. Alternative Markov properties have been proposed for chain graphs, in particular by Andersson et al. [5] (AMP chain graphs). The LWF interpretation has an intuitive equilibrium interpretation and it factorises according to the graph, whereas AMP chain graphs only partially factorise. The complete factorisation is useful in this context as this allows the qualitative representation of some of its factors, so we concern ourselves with LWF chain graphs only. The results of this chapter do not generalise to the other chain graph interpretations. The concepts introduced here are in accordance with existing literature on probabilistic graphical models [32] and chain graphs [197]. They are illustrated by an example chain graph G' as shown in Figure 8.4.

GRAPHICAL CONCEPTS Let $G = (V, E)$ be a *hybrid graph*, where V denotes the set of *vertices* and E the set of *edges*, where an edge is either directed, also called an *arc* , or undirected, also called a *line*. Let indexed letters, e.g., V_1 and V_2, indicate vertices of a chain graph. We denote an arc connecting two vertices by '$\rightarrow$' and a line by '$-$'. Suppose two vertices V_1 and V_2 are in E. If $V_1 \rightarrow V_2$ then V_1 is a *parent* of V_2. If $V_1 - V_2$ then V_1 is a *neighbour* of V_2. The set of parents and neighbours of a vertex V_i are denoted by $\mathrm{pa}(V_i)$ and $\mathrm{ne}(V_i)$, respectively. The set $\mathrm{pa}(V_i) \cup \mathrm{ne}(V_i)$ is the *boundary* of V_i, denoted by $\mathrm{bd}(V_i)$. For example, in G', it holds that $\mathrm{bd}(D) = \{C, E\}$. We will denote $\mathrm{cl}(V_i)$ as the *closure* of V_i defined by $\mathrm{bd}(V_i) \cup \{V_i\}$. These concepts are also assumed to be defined over sets of variables, e.g., if $W \subseteq V$, then $\mathrm{bd}(W) = \bigcup_{V_i \in W} \mathrm{bd}(V_i)$.

One of the key concepts in chain graphs is a *route*, which is a sequence of vertices $V_1, \ldots, V_{n+1}$, such that $V_i - V_{i+1} \in E$, $V_i \rightarrow V_{i+1} \in E$, or $V_i \leftarrow V_{i+1} \in E$. This concept is distinct from a *path*, which is a *route* where no vertex appears more than once. For example, in G', the sequence $A \rightarrow B - C \rightarrow H \leftarrow C$ is a route, but not a path. By a *section* of a route ρ, we mean a maximal undirected subroute $\sigma \subseteq \rho$: $V_i - \ldots - V_j$ with $1 \leqslant i \leqslant j \leqslant n+1$. Note that a section can consist of a single vertex, e.g., in the route $B - C \rightarrow I \leftarrow E$, the sections are $\{B - C, I, E\}$. A section $V_i - \ldots - V_j$ is called a *head-to-head* section on a route ρ if $V_{i-1} \rightarrow V_i$ and $V_j \leftarrow V_{j+1}$ are on the

route ρ. For example, in a route $A \to B - C - D \leftarrow E$, the section $B - C - D$ is called a head-to-head section. A *cycle* is a route (with $n > 1$) where the first and last vertex are the same. A *descending route* is a route, where there are no $V_i \leftarrow V_{i+1} \in E$. A vertex V_i is an ancestor of V_j if there exists a descending route from V_i to V_j. The set $an(V_i)$ denotes the set of *ancestors* of V_i. A *directed route* is a route which includes at least one arc, and where all arcs have the same direction. We will call a vertex V_i a *predecessor* of V_j if there exists a directed route from V_i to V_j. Notice the difference between ancestors and predecessor: if a vertex is a predecessor of another vertex, it is also an ancestor of that vertex, but not vice versa. For example, C has five ancestors in G' (A, B, C, D, and E), but only two predecessors (A and E). C is an ancestor of itself because of, for example, the route $C - B - C$.

A *chain graph* is a hybrid graph with the restriction that no directed cycles, i.e., a directed route which is a cycle, exist. Removing all the arcs from the graph leaves us with vertices connected by lines, called *chain components*; the set of all chain components is denoted here by $\mathcal{C}$. For example, G' contains five chain components: $\{A\}$, $\{E\}$, $\{B, C, D\}$, $\{H\}$, and $\{I\}$. A chain component with its parents plays an important role in the factorisation as shown below. For this reason, we introduce a final graphical concept which is the *family* of a vertex V_i, denoted by $fa(V_i)$, as the set $C \cup pa(C)$ where $C \in \mathcal{C}$ and $V_i \in C$.

GLOBAL MARKOV PROPERTY There exists a simple separation criterion for reading off independence statements from a chain graph, which was introduced by Studený and Bouckaert [197]. Having a route ρ and a set of vertices $Z \subset V$, we say that ρ is *hit* by Z if a vertex of ρ belongs to Z. If ρ is not hit by Z, it will be called *free* with respect to Z. A route is called *superactive* if for every section σ of ρ, the section σ is hit by Z iff σ is a head-to-head section w.r.t. ρ. A set of vertices X (in the graph G) is *c-separated* from a set of vertices Y by the set of vertices Z, denoted by $X \perp\!\!\!\perp Y \mid Z$, if there are no superactive routes between X and Y. Equivalently, we can say that X and Y are c-separated given Z if for every route ρ in G between X and Y, there exists a section σ of ρ such that:

- either σ is a head-to-head section w.r.t. ρ, and σ is free w.r.t. Z;
- or σ is not a head-to-head section w.r.t. ρ, and σ is hit by Z.

If X is *not* c-separated from Y given Z, then this is denoted by $X \not\perp\!\!\!\perp Y \mid Z$. Consider again the graph G'. If $Z = \{H\}$, then the route $A \to B - C - D \leftarrow E$ is not superactive as the head-to-head section $B - C - D$ is free w.r.t. Z. However, the route $A \to B - C \to H \leftarrow C - D \leftarrow E$ is superactive as the only head-to-head section (H) is hit by Z, and all the other sections are free w.r.t. Z. This implies that $A \not\perp\!\!\!\perp E \mid H$.

FACTORISATION Associated to a chain graph $G = (V, E)$ is a joint probability distribution over the set of vertices V that is faithful to the chain graph G, i.e., it contains all the independences implied by the global Markov property. Such distributions can be factorised by an *outer factorisation*:

$$P(V) = \prod_{C \in \mathcal{C}} P(C \mid pa(C)) \tag{8}$$

with $V = \bigcup_{C \in \mathcal{C}} C$, and where each $P(C \mid pa(C))$ is defined by a clique-wise factorisation:

$$P(C \mid pa(C)) = Z^{-1}(pa(C)) \prod_{M \in M_C} \varphi_M(M) \tag{9}$$

given that M_C are the complete (fully connected) subsets in the *closure* graph of C, i.e., the subgraph $G_{C \cup pa(C)}$ where each arc is replaced by a line and each pair of vertices of $pa(C)$ is also connected by a line, also referred as to moralisation. The functions φ are non-negative real functions, called *potentials*; they generalise joint probability distributions in the sense that they do not need to be normalised. Finally, the normalising factor Z is defined as:

$$Z(pa(C)) = \sum_{C} \prod_{M \in M_C} \varphi_M(M) \tag{10}$$

Conversely, (discrete) distributions that factorise in this way are almost always (in a measure-theoretic sense) faithful to the graph G [153].

As a Bayesian network is a special case of a chain graph model where each chain component consists of a single vertex, the *chain graph Markov property* simplifies in that case to

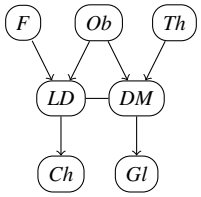
$$P(V) = \prod_{V_i \in V} P(V_i \mid pa(V_i))$$

which is the well-known factorisation theorem of Bayesian networks [32]. In this case, the chain components are formed by single random variables. Therefore, for each of those random variables the distribution is defined as the conditional probability function of this variable, given the value of its parents.

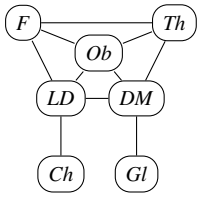
INTERPRETATION Undirected edges in chain graphs can be interpreted as an equilibrium (steady-state) in a feedback model [107]. For example, consider again the graph in Figure 8.3, where there is a feedback relationship between lipid disorder and diabetes mellitus. In practice, this feedback system is in a steady state, although the setpoint of the feedback system may be changed, for example by the amount of insulin given in the therapy. Therefore, only the relationships between variables within a steady-state are relevant, rather than the underlying dynamic process that leads to the equilibrium. Moreover, the underlying dynamics are very difficult to measure *in vivo*, hence, the parameters of such dynamical models, e.g., ordinary differential equations, are difficult to elicit. Therefore, we argue that chain graphs offer an attractive abstraction of the underlying dynamic mechanism for feedback systems in disease models, as there is plenty of data derived from patients in a state of near-equilibrium, i.e., homoeostasis. The corresponding chain graph and factorisation of Figure 8.3 in its steady state is shown in Figure 8.5.

8.3.3 *QPNs*

Qualitative probabilistic networks (QPNs) were introduced by [221], as a qualitative abstraction of Bayesian networks. Conditional probability distributions are replaced by



(i)



(ii)

$$\begin{aligned} P(F, Ob, Th, LD, DM, Ch, Gl) = \\ P(Ch \mid LD) \cdot P(Gl \mid DM) \cdot \\ \varphi_1(F, Ob, LD) \cdot \varphi_2(Ob, LD, DM) \cdot \\ \varphi_3(F, Ob, Th) \cdot \varphi_4(Ob, Th, DM) \cdot \\ P(Ob) \cdot P(Th) \cdot P(F) \end{aligned}$$

(iii)

Figure 8.5: Chain graph representation (i), the closure graph of chain components (ii), and the factorisation (iii) of the example in Figure 3.

qualitative knowledge in the form of signs, which describe the relationships among variables by the concepts of probabilistic influences and synergies. Here we briefly recall the theory in accordance with the definitions in [221, 164, 117].

A *qualitative influence* expresses how the value of one variable influences the probability of observing values of another variable. Let Z denote the set of variables $\mathrm{pa}(B) - A$. We say that A has a *positive qualitative influence on* B, written as $S^+(A, B)$, if

$$P(b \mid a, Z) \geqslant P(b \mid \bar{a}, Z)$$

A *negative* influence, written as $S^-(A, B)$, and a *zero* influence, written as $S^0(A, B)$, are defined analogously, by replacing $\geqslant$ with $\leqslant$ and $=$ respectively. Finally, it always holds that an influence is *ambiguous*, written as $S^?(A, B)$, in particular if none of the other cases hold. Note that all types are influences are mutually consistent, e.g., if both $S^-(A, B)$ and $S^+(A, B)$ holds, then this implies $S^0(A, B)$.

Influences adhere to a set of convenient properties [221]. The property of *symmetry* guarantees that if an influence, say $S^+(A, B)$, exists, the influence $S^+(B, A)$ also exists. Qualitative influences are *transitive*, i.e., the qualitative influences along a directed path (as defined in Section 8.3.2) between two variables, specifying at most one incoming arc for each variable, combine into a single influence using the $\otimes$ operator from Table 8.1. The property of *composition* further asserts that multiple qualitative influences between two variables along parallel paths combine into a single influence between these variables using the $\oplus$ operator from Table 8.1.

In addition to influences, a qualitative probabilistic network includes synergies modelling interactions between influences. An *additive synergy* expresses how the interaction between two variables influences the probability of observing the values of a third variable. Now, let Z denote the set consisting of the variables $\mathrm{pa}(B) - A_1 A_2$. We say there is a *positive additive synergy* of A_1 and A_2 on B, written as $Y^+(\{A_1, A_2\}, B)$, if

$$P(b \mid a_1, a_2, Z) + P(b \mid \bar{a}_1, \bar{a}_2, Z) \geqslant P(b \mid \bar{a}_1, a_2, Z) + P(b \mid a_1, \bar{a}_2, Z)$$

A *product synergy* is used to provide *intercausal reasoning*, i.e., it expresses how upon observation of a common child of two vertices, observing the value of one parent vertex influences the probability of observing a value of the other parent. We say there is a *positive product synergy* of A_1 and A_2 with regard to the value b of variable B, written as $X^+(\{A_1, A_2\}, b)$, if

$$P(b \mid a_1, a_2, Z) \cdot P(b \mid \bar{a}_1, \bar{a}_2, Z) \geqslant P(b \mid \bar{a}_1, a_2, Z) \cdot P(b \mid a_1, \bar{a}_2, Z)$$

	+	-	0	?
+	+	-	0	?
-	-	+	0	?
0	0	0	0	0
?	?	?	0	?

(a) Sign product ($\otimes$)

	+	-	0	?
+	+	?	+	?
-	?	-	-	?
0	+	-	0	?
?	?	?	?	?

(b) Sign sum ($\oplus$)

Table 8.1: Operations on signs.

Negative, *zero*, and *ambiguous* additive and product synergies are defined analogously. It has been shown that the sign of the product synergy implies the sign of the influence between causes given the observation of the child [76, 222], for a given Z. Therefore, the product synergy expresses *intercausal reasoning* to some extent [47].

8.4 QUALITATIVE CHAIN GRAPHS

In this section, we will analyse influences and synergies in the context of chain graph models. The resulting representation will be referred to as *qualitative chain graphs* (QCGs).

8.4.1 *Influences in chain graphs*

The properties of signs in qualitative probabilistic networks rely on the fact that signs hold in any context, i.e., intuitively, a variable A positively influences another variable B if in any possible context the probability of b is higher for a compared to $\bar{a}$. While such a context is relatively clear in case of directed arcs, it is more subtle for probabilistic chain graphs, in which influences can also exist through lines. From a technical point of view, multiple definitions of influences are possible, e.g., the context of the influence may be defined based on e.g., the parents, the neighbours, or possibly the parents of a chain component. Besides technical considerations, a proper definition is also particularly relevant for knowledge elicitation, as the domain expert has to be able to understand influences without reference to the technical details. We believe that such a natural semantics can be defined by means of defining influences in terms of *interventions* [107] on particular variables in the chain graph. For most domain experts, the effect of one variable after the intervention on another variable is an easy to understand concept and is captured in the following formal definition.

Definition 8. *The influence of* A *on* B *in a context* $c \in V - AB$*, where* A *and* B *are two vertices, is the probability* $P(b \parallel a, c) - P(b \parallel \bar{a}, c)$ *where* $P(B \parallel A, C)$ *is the probability of* B *after an intervention on* A *and* C*.*

In other words, the influence of A on B, in a particular context, can be defined as the difference in probability of B from a situation where A is manipulated to $A = \textit{true}$ to a situation where A is manipulated to $A = \textit{false}$. For example, obesity increases

the chance of a lipid disorder (see Example 3); conversely, it is not thought that lipid disorder influences the weight of the patient, even though lipid disorder has a predictive value for 'Obesity'.

It is well-known that without any information about causality, the distribution after an intervention cannot be established. However, similarly to directed graphs [150], a causal interpretation can be given to chain graphs. Given this causal chain graph interpretation, where chain components are interpreted as equilibria, influences between random variables can be restated in terms of conditional probabilities. This forms the basis of the qualitative chain graph theory that will be developed in the remainder of this chapter. In the following lemma, we will first relate the marginal probability of a random variable after an intervention to a conditional probability. This causal interpretation of chain graphs will be assumed throughout the rest of the chapter.

Lemma 2. *If a chain graph* $G = (V, E)$ *is generated by a causal feedback model where lines represent equilibria [107], and* $P(B \parallel V - B)$ *denotes the probability distribution of* B *after an intervention on all other variables, then:*

$$P(B \parallel V - B) = P(B \mid fa(B) - B)$$

Proof. This is a direct corollary of Equation (18) in [107], which states that:

$$P(V - A \parallel A) = \prod_{C \in \mathcal{C}} P(C - A \mid pa(C), C \cap A)$$

Now let $A = V - B$. Given some chain component $C \in \mathcal{C}$, suppose that $B \notin C$, then $P(C - (V - B) \mid pa(C), C \cap (V - B)) = 1$, because $C - (V - B) = \varnothing$. Let $C' \in \mathcal{C}$ be the chain component with $B \in C'$. It then follows that:

$$P(B \parallel V - B) = P(B \mid pa(C'), C' \cap (V - B)) = P(B \mid pa(C'), C' - B)$$

Finally, note that $pa(C') \cup (C' - B) = fa(B) - B$. □

From this property it is straightforward to show that one can generalise the definition of qualitative influences. In order to do so, a convenient tool is the local Markov property of chain graphs, which states that a vertex is independent of its ancestors that are not in its closure given its boundary. This implies the following lemma.

Lemma 3. *Given a chain graph* $G = (V, E)$ *and vertices* $A, B \in V$ *such that* $A \in bd(B)$*, then the following three families of conditional probability distributions are equivalent:*

1. $P(B \mid an(B) - B)$
2. $P(B \mid fa(B) - B)$
3. $P(B \mid bd(B))$

Proof. Observe that:

$$bd(B) \subseteq fa(B) - B \subseteq an(B) - B$$

so in each case, B is conditioned on $\mathrm{bd}(B)$. Furthermore, it holds that the remaining conditioning variables are related to the set of ancestors of B as follows:

$$\begin{aligned} \mathrm{fa}(B) - B - \mathrm{bd}(B) &\subseteq \mathrm{an}(B) - \mathrm{cl}(B) \\ \mathrm{an}(B) - B - \mathrm{bd}(B) &= \mathrm{an}(B) - \mathrm{cl}(B) \end{aligned}$$

Since the local Markov property of chain graph states [57]:

$$V_i \perp\!\!\!\perp \mathrm{an}(V_i) - \mathrm{cl}(V_i) \mid \mathrm{bd}(V_i)$$

the equalities follow. □

Then, using these two lemmas, we obtain an expression of influences in chain graphs in terms of conditional probabilities.

Proposition 2. *Given a chain graph* $G = (V, E)$ *and vertices* $A, B \in V$ *such that* $A \in \mathrm{bd}(B)$*, and a context* c*. Let* $Z = \mathrm{bd}(B) - A$ *and* z *the variable* Z *instantiated according to* c*, i.e.,* $c \equiv z \wedge x$*, where* $X = V - ZAB$*, then the influence of* A *on* B *in context* c *– as formalised in terms of interventions in Definition 8 – equals:*

$$P(b \mid a, z) - P(b \mid \bar{a}, z)$$

Proof. By Definition 8, the influence in context c is:

$$P(b \parallel a, c) - P(b \parallel \bar{a}, c)$$

By Lemma 2, this is equivalent to:

$$P(b \mid a, \mathrm{fa}(B) - AB) - P(b \mid \bar{a}, \mathrm{fa}(B) - AB)$$

Finally, by Lemma 3, this is equivalent to:

$$P(b \mid a, z) - P(b \mid \bar{a}, z)$$

□

Note that it follows that the influence of A on B is a zero influence if $A \notin \mathrm{bd}(B)$ and $A \in \mathrm{an}(B)$. Also note that the qualitative influences generalise the QPN definitions, since $\mathrm{bd}(B) = \mathrm{pa}(B)$ for any $B \in V$ if every chain component consists of a single vertex.

8.4.2 *Qualitative influences*

Given the properties of influences in chain graphs, we are now in the position to define the usual notions of qualitative probabilistic networks for chain graphs. First, we will define qualitative influences, starting with positive qualitative influences, which is defined as a positive influence in all possible contexts.

Definition 9. *Given a chain graph* $G = (V, E)$ *with* $A, B \in V$*, such that* $A \in \mathrm{bd}(B)$*, then vertex* A *positively influences a vertex* B*, written as* $S^+(A, B)$*, if*

$$P(b \mid a, \mathrm{bd}(B) - A) \geqslant P(b \mid \bar{a}, \mathrm{bd}(B) - A)$$

for all configurations of $\mathrm{bd}(B) - A$*.*

Generally, if $\delta \in \{+, -, 0, ?\}$ then $S^{\delta}(A, B)$ denotes the qualitative influence between A and B. The negative (-), zero (0), and ambiguous (?) influences are defined in line with QPNs, i.e., for the negative and zero influence we replace $\geqslant$ by $\leqslant$ and $=$, respectively. Finally, $S^{?}(A, B)$ always holds if $A \in \mathrm{bd}(B)$.

Qualitative influences in QPNs adhere to the properties of symmetry, transitivity, and composition [221, 76], which form the basis for qualitative inference in QPNs. Symmetric means that if there is some influence from a vertex A to a vertex B, then there is an influence from B to A with the same sign if the arc is reversed.[2] Therefore, only a single sign is needed for every arc in a QPN. In the following, we will prove that this symmetry is preserved for qualitative chain graphs, i.e., also for neighbouring vertices the signs are symmetric. First we prove a lemma that rephrases qualitative influences in terms of relationships between potential functions. We will focus in this lemma and theorem on positive influences, however, the same reasoning holds for negative and zero influences.

Lemma 4. *Given a chain graph* G *containing vertices* A *and* B*, with* $A \in \mathrm{bd}(B)$ *and* B *an element of a component* C*, it holds that:*

$$P(b \mid a, \mathrm{fa}(B) - AB) \geqslant P(b \mid \bar{a}, \mathrm{fa}(B) - AB)$$

if and only if

$$\prod_{M \in M_{AB}} \varphi_M(a, b)\varphi_M(\bar{a}, \bar{b}) \geqslant \prod_{M \in M_{AB}} \varphi_M(a, \bar{b})\varphi_M(\bar{a}, b)$$

where $M_{AB} = \{M \in M_C \mid \{A, B\} \subseteq M\}$ *and* $\varphi_M(A, B)$ *is shorthand for* $\varphi(M - AB, A, B)$.

Proof. In the following, we will write $\varphi_M(a, X)$ for $\varphi_M(X)$ if $A \notin M$ for any $X \subseteq V$. This allows us to consider all cliques conditioned on a certain variable (e.g. A) without making a distinction between those that contain A and those that do not contain A. By basic probability theory, we have:

$$P(B \mid A, \mathrm{fa}(B) - AB) = \frac{P(C, \mathrm{pa}(C))}{P((C, \mathrm{pa}(C)) - B)} = \frac{P(C \mid \mathrm{pa}(C))P(\mathrm{pa}(C))}{\sum_B P(C \mid \mathrm{pa}(C))P(\mathrm{pa}(C))}$$

Using Equation (9), that factorises conditional probabilities of a component into potentials, the left-hand side therefore equals to:

$$\frac{Z^{-1}(\mathrm{pa}(C))\Big(\prod_{M_C} \varphi_M(a, b)\Big)P(\mathrm{pa}(C))}{\sum_B Z^{-1}(\mathrm{pa}(C))\Big(\prod_{M_C} \varphi_M(a, B)\Big)P(\mathrm{pa}(C))} \geqslant \frac{Z^{-1}(\mathrm{pa}(C))\prod_{M_C}\Big(\varphi_M(\bar{a}, b)\Big)P(\mathrm{pa}(C))}{\sum_B Z^{-1}(\mathrm{pa}(C))\Big(\prod_{M_C} \varphi_M(\bar{a}, B)\Big)P(\mathrm{pa}(C))}$$

2 Note that parents are added during arc reversals in Bayesian networks. See [221] for details.

Given that $B \in C$, we have $B \notin pa(C)$, so the term $Z^{-1}(pa(C))P(pa(C))$ only depends on A. By replacing this term by $f(A)$ and multiplying each side by the denominators, we obtain:

$$\prod_{M_C} \varphi_M(a,b)f(a) \sum_B \prod_{M_C} \varphi_M(\bar{a},B)f(\bar{a}) \geqslant \prod_{M_C} \varphi_M(\bar{a},b)f(\bar{a}) \sum_B \prod_{M_C} \varphi_M(a,B)f(a)$$

Writing out the possible values for the summation over B, i.e., b and $\bar{b}$, we obtain:

$$\prod_{M_C} \varphi_M(a,b)\varphi_M(\bar{a},b)f(a)f(\bar{a}) + \prod_{M_C} \varphi_M(a,b)\varphi_M(\bar{a},\bar{b})f(a)f(\bar{a}) \geqslant \prod_{M_C} \varphi_M(\bar{a},b)\varphi_M(a,b)f(a)f(\bar{a}) + \prod_{M_C} \varphi_M(\bar{a},b)\varphi_M(a,\bar{b})f(a)f(\bar{a})$$

Removing the factors that are the same on both sides of the equation we get:

$$\prod_{M_C} \varphi_M(a,b)\varphi_M(\bar{a},\bar{b}) \geqslant \prod_{M_C} \varphi_M(a,\bar{b})\varphi_M(\bar{a},b)$$

For all potentials not depending on both A and B, e.g., $\varphi_M(A,B) = \varphi_M(B)$, its corresponding factors are also the same on both sides of the equation, leaving us with:

$$\prod_{M \in M_{AB}} \varphi_M(a,b)\varphi_M(\bar{a},\bar{b}) \geqslant \prod_{M \in M_{AB}} \varphi_M(a,\bar{b})\varphi_M(\bar{a},b)$$

□

Example 4. *Continuing Example 3, to evaluate a (say, positive) influence of Ob on LD only involves φ_1 and φ_2 (cf. Figure 8.5(*iii*)). Therefore, a positive influence of Ob on LD is equivalent to:*

$$\varphi_1(F,ob,ld)\ \varphi_1(F,\overline{ob},\overline{ld})\ \varphi_2(ob,ld,DM)\ \varphi_2(\overline{ob},\overline{ld},DM) \geqslant \varphi_1(F,\overline{ob},ld)\ \varphi_1(F,\overline{ob},ld)\ \varphi_2(\overline{ob},ld,DM)\ \varphi_2(ob,\overline{ld},DM)$$

for all values of F and DM.

In other words, determining the nature of a qualitative influence between two vertices implies that one only has to consider those potentials for cliques containing the two variables that describe the influence. In general, we have the following result that we were aiming for, proving the symmetry property of qualitative influences.

Theorem 2. *It holds that qualitative signs of chain graphs are symmetric, i.e., given a chain graph $G = (V,E)$, suppose $(A,B) \in E$ (i.e. $A - B$ or $A \to B$), then $P(b \mid a,X) - P(b \mid \bar{a},X) \geqslant 0$ if and only if $P(a \mid b,X,Y) - P(a \mid \bar{b},X,Y) \geqslant 0$, where $X = bd(B) - A$ and $Y = bd(A) - B$.*

Proof. Suppose $P(b \mid a, X) \geqslant P(b \mid \bar{a}, X)$. Assume $P(b \mid a, X) - P(b \mid \bar{a}, X) \geqslant 0$. Since $Y \subseteq \mathrm{an}(B) - \mathrm{cl}(B)$, by the local Markov property, we have $B \perp\!\!\!\perp Y \mid AX$, so it follows that $P(b \mid a, X, Y) - P(b \mid \bar{a}, X, Y) \geqslant 0$. Let $Z = X \cup Y$. Then, by Bayes' rule, we have:

$$P(b \mid a, Z) \geqslant P(b \mid \bar{a}, Z) \quad \Leftrightarrow \quad \frac{P(a \mid b, Z)P(b, Z)}{P(a, Z)} \geqslant \frac{P(\bar{a} \mid b, Z)P(b, Z)}{P(\bar{a}, Z)}$$

So then this is equivalent to:

$$\begin{aligned}
& \frac{P(a \mid b, Z)}{\sum_B P(a \mid B, Z) \mid P(B, Z)} \geqslant \frac{P(\bar{a} \mid b, Z)}{\sum_B P(\bar{a} \mid B, Z)P(B, Z)} \\
\Leftrightarrow \quad & P(a \mid b, Z) \sum_B P(\bar{a} \mid B, Z)P(B, Z) \\
& \quad \geqslant P(\bar{a} \mid b, Z) \sum_B P(a \mid B, Z) \mid P(B, Z) \\
\Leftrightarrow \quad & P(a \mid b, Z)P(\bar{a} \mid \bar{b}, Z) \geqslant P(\bar{a} \mid b, Z)P(a \mid \bar{b}, Z) \\
\Leftrightarrow \quad & P(a \mid b, Z)(1 - P(a \mid \bar{b}, Z)) \geqslant P(a \mid \bar{b}, Z)(1 - P(a \mid b, Z)) \\
\Leftrightarrow \quad & P(a \mid b, Z) \geqslant P(a \mid \bar{b}, Z)
\end{aligned}$$

□

Note that this does not prove $S^+(A, B)$ iff $S^+(B, A)$ in a strict sense, as an influence of B on A is only defined if B is in the boundary of A. Instead it should be seen as the influence after the reversal of the arc where the A obtains the boundary of B, which is similar to arc reversal in Bayesian networks and QPNs [188, 221].

As mentioned, besides symmetry, there are two other important properties. First, for a single route from vertices A to B in a chain graph G, the influences along a path are *transitive*, and the influence of A on B is the *sign-product* (Table 8.1a) of all influences along the path. For multiple *paths* between two vertices A and B, the influences of these routes are *composite*, i.e., the influence of A on B is the *sign-sum* (Table 8.1b) of all influences of each paths. These properties are combined in the following theorem.

Theorem 3. *Let* $S^{\delta}(A, B, G)$ *be the qualitative influence in the distribution of chain graph* G*, and let* $\mathrm{red}(B, G)$ *be the probability distribution of chain graph* G *with* B *marginalised out. It holds that* $S^{\delta_1}(A, B, G) \wedge S^{\delta_2}(B, C, G) \wedge S^{\delta_3}(A, C, G) \Rightarrow S^{(\delta_1 \otimes \delta_2) \oplus \delta_3}(A, C, \mathrm{red}(B, G))$.

Proof. Let $X = \mathrm{bd}(B) - A$ and $Y = \mathrm{bd}(C) - AB$. Observe that:

$$\begin{aligned}
P(C \mid A, X, Y) &= \sum_B P(C, B \mid A, X, Y) \\
&= \sum_B P(C \mid A, B, X, Y)P(B \mid A, X, Y)
\end{aligned}$$

Choose an arbitrary X and Y, and let $\Delta_{AB} = P(b \mid a, XY) - P(b \mid \bar{a}, XY)$, $\Delta_{BCa} = P(c \mid a, b, XY) - P(c \mid a, \bar{b}, XY)$, $\Delta_{ACb} = P(c \mid a, b, XY) - P(c \mid \bar{a}, b, XY)$, and

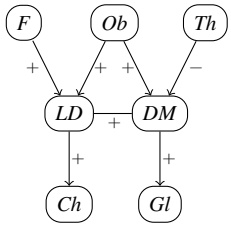


Figure 8.6: Qualitative representation with influences of the chain graph from Figure 8.3.

$\Delta_{AC\bar{b}} = P(c \mid a, \bar{b}, XY) - P(c \mid \bar{a}, \bar{b}, XY)$. Furthermore, define $x = P(c \mid a, b, XY)$ and $y = P(b \mid a, XY)$. Then:

$$\begin{aligned} P(c \mid a, XY) - P(c \mid \bar{a}, XY) &= xy + (x - \Delta_{BCa})(1 - y) \\ &\quad -(x - \Delta_{ACb})(y - \Delta_{AB}) \\ &\quad -(x - \Delta_{AC\bar{b}} - \Delta_{BCa})(1 - y + \Delta_{AB}) \\ &= \Delta_{AC\bar{b}}((1 - y) + \Delta_{AB}) + \Delta_{ACb}(y - \Delta_{AB}) \\ &\quad +\Delta_{AB}\Delta_{BCa} \end{aligned}$$

Note that $1 - y + \Delta_{AB} = P(\bar{b} \mid \bar{a}, XY) \geqslant 0$ and $y - \Delta_{AB} = P(b \mid \bar{a}, XY) \geqslant 0$, so it follows that:

$$\text{sign}(P(c \mid a, XY) - P(c \mid \bar{a}, XY)) = \text{sign}(\Delta_{AC\bar{b}} + \Delta_{ACb} + \Delta_{AB}\Delta_{BCa})$$

The properties can then be verified. Note that both $\Delta_{AC\bar{b}}$ and Δ_{ACb} have the sign δ_3; Δ_{AB} has the sign δ_1; Δ_{BCa} has the sign δ_2. The sign-product operator behaves as a product and the sign-sum operator behaves as a sum. Note that in case $\delta_3 = 0$, we have transitivity only. □

Example 5. *Consider Figure 8.6, which is a qualitative version of the model in Figure 8.3. Since* $S^+(Ob, DM)$ *and* $S^+(DM, Gl)$*, we derive, for example,* $S^+(Ob, Gl)$*. Similarly, we can derive* $S^-(Th, Gl)$*, i.e., therapy lowers the chance on a high blood glucose.*

8.4.3 *Additive synergies*

As mentioned in the preliminaries, an *additive synergy* expresses information about the joint influence of two vertices A_1 and A_2 on a neighbour of these vertices. The intuitive idea is that there is a positive synergy if the joint influence is larger than their separate influence on this neighbour. Following Definition 9, we define a synergy within a certain context as follows.

Definition 10. *Given a chain graph* $G = (V, E)$ *and three vertices* A_1, A_2 *and* B*, with* $A_1, A_2 \in \text{bd}(B)$*. Let* c *be any context* $c \in V - A_1B$*. Let* k *be the influence of* A_1 *on* B *in context* $c \wedge a_2$ *and* l *be the influence of* A_1 *on* B *in context* $c \wedge \bar{a}_2$*. Then the synergy between* A_1 *and* A_2 *on* B *in context* c *is* $k - l$*.*

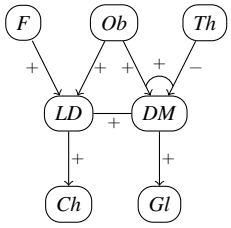


Figure 8.7: The model presented in Figure 8.6 with an additional additive synergy on *DM*.

From Proposition 2, where we show that the influence of A_1 on B in context $c \wedge A_2$ is equal to $P(b \mid a_1, A_2, z) - P(b \mid \bar{a}_1, A_2, z)$ where z is a configuration of $bd(B) - A_1A_2$, it follows that the synergy of A_1 and A_2 on B in context c is positive if $P(b \mid a_1, a_2, z) - P(b \mid \bar{a}_1, a_2, z) \geqslant P(b \mid a_1, \bar{a}_2, z) - P(b \mid \bar{a}_1, \bar{a}_2, z)$. Therefore, again, a probabilistic definition for additive synergies in QCGs naturally follows, which is defined as a synergy in any possible context.

Definition 11. *We say that vertices A_1 and A_2 express a positive additive synergy on a vertex B, written as $Y^+(\{A_1, A_2\}, B)$, iff $A_1, A_2 \in bd(B)$, $Z = bd(B) - A_1A_2$, and*

$$P(b \mid a_1, a_2, Z) - P(b \mid \bar{a}_1, a_2, Z) \geqslant P(b \mid a_1, \bar{a}_2, Z) - P(b \mid \bar{a}_1, \bar{a}_2, Z)$$

Again, the negative and zero synergies are defined similarly by replacing $\geqslant$ with $\leqslant$ or $=$ respectively. Ambiguous synergies always hold.

Example 6. *Continuing Example 5, consider Figure 8.7. In this model, we express a positive additive synergy between the variables obesity and therapy on diabetes mellitus, i.e., $Y^+(\{Ob, Th\}, DM)$. This implies that e.g. the positive influence of therapy on diabetes is positively influenced by weight loss. Conversely, the positive influence of weight-loss on diabetes is positively influenced by treatment.*

It can be verified that additive synergies in chain graphs are similar to additive synergies in QPNs. For example, it is obvious that additive synergies are symmetric, i.e., if $Y^\delta(\{A_1, A_2\}, B)$ then $Y^\delta(\{A_2, A_1\}, B)$. Other properties of synergies have been well studied for QPNs, in particular in context of variable elimination [221]. We do not further study additive synergies in this context as in this chapter we mainly focus on the use of these kind of synergies for imposing constraints on the probability distribution. It seems reasonable to conjecture that similar results to the QPN case will apply.

8.4.4 *Intercausal reasoning and product synergies*

In QPNs, a *product synergy* expresses how the value of one cause influences the probability of the value of another cause when observing the common child. A negative product synergy of A_1 and A_2 given b expresses that if a_1 is the case, then this renders a_2 less likely. While in a QPN a cause is simply the parent of a vertex, it is less clear what it means in the undirected case. First, we will define product synergies in general and then study its properties.

Definition 12. *We say that vertices A_1 and A_2 express a negative product synergy with regard to the value w for the set of vertices W, denoted by $X^-(\{A_1, A_2\}, w)$, iff*

$$P(w \mid a_1, a_2, Z)P(w \mid \bar{a}_1, \bar{a}_2, Z) \leqslant P(w \mid a_1, \bar{a}_2, Z)P(w \mid \bar{a}_1, a_2, Z)$$

where $Z = \mathrm{bd}(W) - A_1A_2$.

Again, the other signs are defined analogously.

As said, the purpose of product synergies is to define intercausal relationships between vertices. First, we will relate intercausal reasoning to product synergies in chain graphs using the following theorem.

Theorem 4. *Given a chain graph $G = (V, E)$ and two vertices $A_1, A_2 \in V$ and a set of vertices $W \subseteq V$. Let $Z = \mathrm{bd}(W) - A_1A_2$, $X = \mathrm{bd}(A_2)$ and suppose that $A_1 \perp\!\!\!\perp A_2 \mid ZX$ and $W \perp\!\!\!\perp X \mid A_1A_2Z$. If we observe w, then we have a negative influence of A_1 on A_2, i.e.,*

$$P(a_2 \mid a_1, w, Z, X) \leqslant P(a_2 \mid \bar{a}_1, w, Z, X)$$

if and only if

$$P(w \mid a_1, a_2, Z)P(w \mid \bar{a}_1, \bar{a}_2, Z) \leqslant P(w \mid \bar{a}_1, a_2, Z)P(w \mid a_1, \bar{a}_2, Z)$$

The same equality holds if we replace $\leqslant$ by $\geqslant$ or $=$.

Proof. By Bayes' theorem we have:

$$P(A_2 \mid A_1, w, Z, X) = \frac{P(w \mid A_1, A_2, Z, X)P(A_2 \mid A_1, Z, X)}{P(w \mid A_1, Z, X)}$$

Since $A_1 \perp\!\!\!\perp A_2 \mid ZX$, the term $P(A_2 \mid A_1, Z, X) = P(A_2 \mid Z, X)$. Similarly, $P(w \mid A_1, A_2, Z, X) = P(w \mid A_1, A_2, Z)$ and $P(w \mid A_1, Z, X) = P(w \mid A_1, Z)$. Now by marginalising over A_2 in the denominator:

$$\begin{aligned} P(A_2 \mid A_1, w, Z, X) &= \frac{P(w \mid A_1, A_2, Z)P(A_2 \mid Z, X)}{\sum_{A_2} P(w, A_2 \mid A_1, Z)} \\ &= \frac{P(w \mid A_1, A_2, Z)P(A_2 \mid Z, X)}{\sum_{A_2} P(w \mid A_1, A_2, Z)P(A_2 \mid Z, X)} \end{aligned}$$

Now we find:

$$\begin{aligned} & P(a_2 \mid a_1, w, Z) \leqslant P(a_2 \mid \bar{a}_1, w, Z) \\ \Leftrightarrow\ & \frac{P(w \mid a_1, a_2, Z)P(a_2 \mid Z, X)}{\sum_{A_2} P(w \mid a_1, A_2, Z)P(A_2 \mid Z, X)} \leqslant \frac{P(w \mid \bar{a}_1, a_2, Z)P(a_2 \mid Z, X)}{\sum_{A_2} P(w \mid \bar{a}_1, A_2, Z)P(A_2 \mid Z, X)} \\ \Leftrightarrow\ & P(w \mid a_1, a_2, Z)P(a_2 \mid Z, X) \sum_{A_2} P(w \mid \bar{a}_1, A_2, Z, X)P(A_2 \mid Z) \leqslant \\ & P(w \mid \bar{a}_1, a_2, Z)P(a_2 \mid Z, X) \sum_{A_2} P(w \mid a_1, A_2, Z)P(A_2 \mid Z, X) \\ \Leftrightarrow\ & P(w \mid a_1, a_2, Z)P(w \mid \bar{a}_1, a_2, Z)P(a_2 \mid Z, X) + \\ & P(w \mid a_1, a_2, Z)P(w \mid \bar{a}_1, \bar{a}_2, Z)P(\bar{a}_2 \mid Z, X) \\ & \leqslant P(w \mid \bar{a}_1, a_2, Z)P(w \mid a_1, a_2, Z)P(a_2 \mid Z, X) + \\ & P(w \mid \bar{a}_1, a_2, Z)P(w \mid a_1, \bar{a}_2, Z)P(\bar{a}_2 \mid Z, X) \\ \Leftrightarrow\ & P(w \mid a_1, a_2, Z)P(w \mid \bar{a}_1, \bar{a}_2, Z) \leqslant P(w \mid \bar{a}_1, a_2, Z)P(w \mid a_1, \bar{a}_2, Z) \end{aligned}$$

□

It is clear that in case we have a QPN, product synergies characterise intercausal reasoning if $A_1, A_2 \in \mathrm{pa}(W)$ and A_1 and A_2 are unconditionally independent. This particular case was discussed in [76] and in subsequent papers. In contrast, in chain graphs, the following holds, which states that in case the main condition of Theorem 4 holds – i.e., that A_1 and A_2 are independent given the boundary of the conditioning vertex and the boundary of A_2 – then these vertices must be predecessors of the conditioning vertices.

Proposition 3. *Given a chain graph* $G = (V, E)$ *and three vertices* $A_1, A_2, B \in V$. *Let* $Z = \mathrm{bd}(B) - A_1A_2$ *and* $X = \mathrm{bd}(A_2)$ *and assume* $A_1 \not\perp\!\!\!\perp B$ *and* $A_2 \not\perp\!\!\!\perp B \mid X$, *and* $B \notin X$. *If* $A_1 \perp\!\!\!\perp A_2 \mid ZX$, *then* A_1 *and* A_2 *are predecessors of* B.

Proof. Suppose not: suppose A_1 is the vertex which is not a predecessor of B (the argument is similar for A_2). Then we have a descending superactive route ρ from B to A_1 w.r.t ZX. Similarly, for A_2, there exists a superactive route σ from A_2 to B w.r.t. ZX. Since $\rho = B - \cdots A_1$ or $\rho = B \to \cdots A_1$, B cannot be a head-to-head section, so, given that $B \notin ZX$, the union of ρ and σ (from A_2 to A_1) is also superactive w.r.t. ZX. This contradicts the assumption $A_1 \perp\!\!\!\perp A_2 \mid ZX$. □

In case we condition on a vertex in a chain component, it follows from the definition of c-separation, that we create a superactive route between parents of a chain component. This could be seen as intercausal reasoning between these parents. For example, if we have a graph $A_1 \to C - B \leftarrow A_2$, there will be an influence of A_1 on A_2 if we condition on B. A natural question is whether there are any relationships between these intercausal influences. In the following theorem we show that they can be composed by the sign-sum operator.

Theorem 5. *Given a chain graph* $G = (V, E)$ *and* $W_1 \subseteq V$ *and* $W_2 \subseteq V$ *two disjoint sets of some chain component, such that* A_1 *and* A_2 *are predecessors of both* W_1 *and* W_2. *If* $X^{\delta_1}(\{A_1, A_2\}, w_1)$ *and* $X^{\delta_2}(\{A_1, A_2\}, w_2)$, *then* $X^{\delta_1 \oplus \delta_2}(\{A_1, A_2\}, w_1 \wedge w_2)$

Proof. Let $Z = \mathrm{bd}(W_1 \cup W_2) - A_1, A_2$, $Z_1 = \mathrm{bd}(W_1) - A_1, A_2$ and $Z_2 = \mathrm{bd}(W_2) - A_1, A_2$. Observe that:

$$P(w_1 \wedge w_2 \mid A_1, A_2, Z) = P(w_1 \mid w_2, A_1, A_2, Z)P(w_2 \mid A_1, A_2, Z)$$

Since $Z_1 \subseteq Z$ and $Z_2 \subseteq Z$, and $W_1 \perp\!\!\!\perp (Z \setminus Z_1) \mid Z_1, A_1, A_2$ and $W_2 \perp\!\!\!\perp (Z \setminus Z_2) \mid Z_2, A_1, A_2$, we find:

$$P(w_1 \wedge w_2 \mid A_1, A_2, Z) = P(w_1 \mid A_1, A_2, Z_1)P(w_2 \mid A_1, A_2, Z_2)$$

Now, suppose for example that δ_1, δ_2 are negative, then given $X^-(\{A_1, A_2\}, w_i)$, for $i \in \{1, 2\}$, Theorem 4 states:

$$P(w_i \mid a_1, a_2, Z_i)P(w_i \mid \bar{a}_1, \bar{a}_2, Z) \leqslant P(w_i \mid a_1, \bar{a}_2, Z_i)P(w_i \mid \bar{a}_1, a_2, Z_i)$$

Then:

$$\begin{aligned}
& P(w_1 \mid a_1, a_2, Z_1)P(w_1 \mid \bar{a}_1, \bar{a}_2, Z_1)P(w_2 \mid a_1, a_2, Z_2)P(w_2 \mid \bar{a}_1, \bar{a}_2, Z_2) \\
\leqslant\ & P(w_1 \mid a_1, \bar{a}_2, Z_1)P(w_1 \mid \bar{a}_1, a_2, Z_1)P(w_2 \mid a_1, \bar{a}_2, Z_2)P(w_2 \mid \bar{a}_1, a_2, Z_2) \\
\Leftrightarrow\ & P(w_1 \wedge w_2 \mid a_1, a_2, Z)P(w_1 \wedge w_2 \mid \bar{a}_1, \bar{a}_2, Z) \\
& \leqslant P(w_1 \wedge w_2 \mid a_1, \bar{a}_2, Z)P(w_1 \wedge w_2 \mid \bar{a}_1, a_2, Z) \\
\Leftrightarrow\ & X^{-}(\{A_1, A_2\}, w_1 \wedge w_2)
\end{aligned}$$

For other pairs of δ_1, δ_2, the result can be verified similarly. □

In practice, this means that if we specify the product synergies $X^{\delta}(\{A_1, A_2\}, b_i)$ and $X^{\delta'}(\{A_1, A_2\}, \bar{b}_i)$ for every B_i in a chain component, where A_1 and A_2 are parents, then for each subset of variables in this chain component, there is a product synergy with a default sign determined by the sign-sum operator. To strengthen this sign, e.g., if the result is an ambiguous sign, intercausal reasoning using subsets in the chain component can be specified.

It would therefore be natural to define all product synergies for pairs of parents of a chain component. However, for QCGs, such intercausal reasoning does not completely coincide with product synergies as we will show in the next proposition.

Proposition 4. *Given a chain graph* $G = (V, E)$ *and three vertices* $A_1, A_2, B \in V$ *such that* A_1 *and* A_2 *are two parents of a chain component* C. *Given any vertex* $B \in C$, *if* $Z = \mathrm{bd}(B) - A_1A_2$, $X = \mathrm{bd}(A_2)$ *and* $A_1 \not\perp\!\!\!\perp A_2 \mid BZX$, *then* A_1 *and* A_2 *are the (direct) parents of* B.

Proof. Suppose without loss of generalisation that A_1 is not a parent of B (though recall that A_1 and A_2 are predecessors). Then there is some $D \in Z$ such that there is a directed route from $\rho = A_1 \to \cdots D - B$. Now take some route σ from B to A_2. Then the union of ρ and σ is not superactive with respect to Z, so $A_1 \perp\!\!\!\perp A_2 \mid BZX$. □

In other words, Theorem 4 can only be applied if there are intercausal relationships between direct parents of a vertex. One might think that this is only because the condition $A_1 \not\perp\!\!\!\perp A_2 \mid ZX$ in this theorem is too strong. While this may be partially true, consider the graph $A_1 \to C - B - D \leftarrow A_2$. Obviously, $P(B \mid A_1, A_2, C, D) = P(B \mid C, D)$. Therefore the product synergy of A_1 and A_2 given b is necessarily zero, even though $A_1 \not\perp\!\!\!\perp A_2 \mid B$. Therefore, it is impossible that product synergies can completely characterise intercausal reasoning in chain graphs. Previously, it was already shown that product synergies do not characterise intercausal reasoning in case there are uninstantiated ancestors [47]. This paper generalises this negative result by observing that product synergies also do not characterise intercausal reasoning within chain components in general.

Example 7. *Continuing Example 6, consider Figure 8.8. We expect a negative product synergy* $X^{-}(\{F, Ob\}, ld)$*: for patients who have a lipid disorder, the observation that the patient is obese renders it less likely that the lipid disorder is caused by a familial hypercholesterolaemia. The same effect could be expected for diabetes since diabetes can be caused by a lipid disorder, therefore, we expect a negative intercausal influence given diabetes as well. However, this property cannot be expressed by a product synergy.*

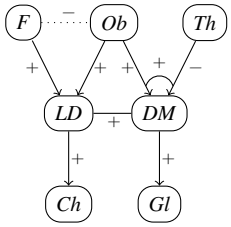


Figure 8.8: Example model with a negative product synergy on *DM*. Graphically, we only specify one, although the product synergies is dependent on the number of variables in the chain component.

8.5 SIGN PROPAGATION

Several papers [45, 214, 165, 110] have studied qualitative inference in QPNs using a qualitative version of belief propagation, which is called *sign propagation*. Similar to belief propagation, the idea is that vertices in the graph maintain a (qualitative) belief and send their beliefs to their neighbours. Upon receiving beliefs from neighbours, beliefs are updated and further distributed. It can be shown that the complexity of sign propagation is linear in the number of vertices in the graph, making it an attractive method for purely qualitative inference.

To generalise sign propagation in chain graphs, we need to introduce some additional concepts, in particular the concept of a *trail* in a CG. However, trails may be confusing as the QPN literature contains at least two notions of trails (cf. [45][3] and [164][4]), which are both distinct from trails in CG [197]. Here we refer to the CG trails, which is a route such that no arc appears twice in this route and each section in the route consists of distinct vertices. A trail can be *active*, which means that none of its sections are blocked by the evidence. This is relevant in this context, because of the following property.

Lemma 5 (Lemma 4.1, [197])**.** *Let* $\mathsf{G} = (\mathsf{V}, \mathsf{E})$ *be a chain graph with vertices* X*,* Y*, and* Z*. Then* $\mathsf{X} \perp\!\!\!\perp \mathsf{Y} \mid \mathsf{Z}$ *iff every trail in* G *from a vertex of* X *to a vertex of* Y *is c-separated by* Z*.*

This shows that it is sufficient to only consider the active trails rather than all routes in the graph. Active trails are also interesting because they directly relate to evidential trails, which are used to propagate signs in QPNs. This shows that the remainder of this section generalises the definitions for QPNs.

Proposition 5. *Given a QPN with and some route* ρ *in the graph. If* ρ *is an evidential trail, which was defined as an active path from* $\mathfrak{e}$ *to* $\mathfrak{n}$ *[45], then it is an active CG trail.*

Proof. Let ρ be an evidential trail. The claim is that $(\mathfrak{i})$ no directed edge appears twice in ρ and $(\mathfrak{ii})$ that each section in the route consists of distinct vertices. Note that each

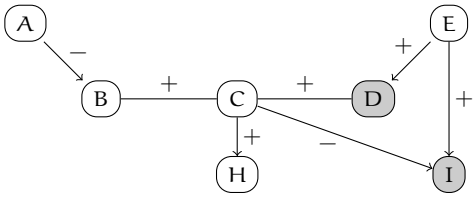
3 In [45], trails correspond to *routes*. In context of Bayesian networks, this can be used to define the concept of d-separation [60].

4 In [164], trails are connected subgraphs between vertices.

section in a QPN consists of a single vertex. Therefore, the condition that vertices appear only once implies (i) and (ii). □

Note that the QPN evidential trail is not a correct definition in the context of chain graphs, because some vertices may appear multiple times in a CG trail. This cannot be ignored as they may make e and n dependent.

Example 8. *Consider the following QCG where* D *and* I *are observed.*



Consider the routes between A *and* H*. There are two active trails, i.e.,* $t_1 = A \to B - C \to H$ *and* $t_2 = A \to B - C - D \leftarrow E \to I \leftarrow C \to H$*. Of course, there are other superactive routes, e.g.,* $A \to B - C - B - C \to H$*. In this case* t_2 *is not an evidential trail, but cannot be ignored: suppose for example that* B *is also observed, then* t_2 *is the only evidential trail between* A *and* H*. Through intercausal reasoning between* A *and* E*, and* C *and* E*, it may be the case that* A *has an effect on* H*.*

In case there is a single evidential trail between two vertices, we can use the transitivity and symmetry property to compute the qualitative signs.

Lemma 6. *Given an active trail* ρ *in a QCG* G *from* A *to* B *with observed vertices* O*. Remove all the head-to-head sections* $X \to \ldots \leftarrow Y$ *in* ρ *and replace them by a line with an influence according to the product synergy in* G *and the observations* O*. If there are no other active trails besides* ρ*, then the sign product of the intermediate influences is the influence between* A *and* B*.*

Proof. (sketch) After removing the head-to-head sections, we have a path between A and B. If there are diverging vertices, these can be reversed using Theorem 2. After this, Theorem 3 can be applied repeatedly to marginalise out the intermediate vertices in this graph. Note that the boundary of the vertex changes during arc reversals and marginalisation; the influence however holds for any context. □

This provides a way to perform sign propagation in singly-connected chain graphs, i.e., chain graphs where the underlying undirected graph is a tree. This is because of the following property of chain graphs.

Proposition 6. *Given a singly-connected chain graph* $G = (V, E)$*, and given a pair of vertices* $A, B \in V$ *and observations* $O \subseteq V$*, if* $A \not\perp\!\!\!\perp B \mid O$*, there exists a single active trail between* A *and* B *given* O*.*

Proof. It is sufficient to show that there is at most one trail between the vertices A and B: if $A \not\perp\!\!\!\perp B \mid O$, then this trail must exist and be active. Suppose there are two trails, then clearly, because the chain graph is singly-connected, one of the trails must contain the same vertex at least twice. Let ρ be this trail, which contains a subtrail σ from X

to X. It cannot be that σ is completely undirected (i.e., they are in the same section), because then ρ is not a trail. But then there must be Y in σ with a route from X to Y that contains a directed edge and a route from Y to X that contains the same edge. This also contradicts the assumption that ρ is a trail. □

For two vetices with multiple evidential trails, sign propagation is difficult. The original algorithm by Druzdzel and Henrion [45] propagates the signs over all trails and then combines signs by a sign-sum operator. For a long time, it was assumed that this algorithm was correct. More recently, van Kouwen et al. [214] showed that this may not always give the correct result in multiply-connected networks. One issue is that direct influences dominate influences through intercausal mechanisms [44]. Another issue is that the direction of arcs in a trail matter when propagating signs. Finally, if there are multiple observations, the result may depend on the order of observations or may be unnecessarily ambiguous [165]. To overcome these problems, solutions have been suggested that solve these issues, resulting in a complex inference algorithm [214, 165, 110]. It should be noted that while these solutions overcome the problems that have been recognised in the original algorithm, as far as we are aware, a general correctness and completeness proof of recent algorithms for sign propagation in QPN is an open problem.

8.6 EXPERIMENTAL RESULTS

Since chain graphs have been shown to model equilibrium systems, as mentioned in the introduction, the theoretical foundation of the previous sections can be used to explore the qualitative dynamics in such systems. Here we continue with the running example that assumes that there is a kind of equilibrium between diabetes mellitus and lipid disorders, and that its state can be influenced by other (patho)physiological conditions. These influences can now be expressed in terms of qualitative influences and synergies, and the resulting QCG can be used to perform probabilistic inference.

While probabilistic inference with a QCG can be done using sign propagation, based on message-passing between neighbouring vertices as presented in the previous section, such inference has serious limitations in case of trade-offs, i.e., when there are two opposite influences. Consider again Figure 8.6 and suppose we need to make a decision for an obese patient. In case the patient loses weight and therapy is started, it is clear that this will reduce the blood sugar (by $\mathrm{S}^{+}(Ob, DM)$ and $\mathrm{S}^{-}(Th, DM)$). Often, however, weight reduction is unsuccessful. In that case, the effect of therapy is unclear; obesity will make diabetes more likely and the therapy makes it less likely; so by the sign-sum operator the whole effect is ambiguous.

An alternative approach is to look upon the qualitative signs as constraints on the joint probability distribution, as proposed in [46], where a canonical representation consisting of (in)equalities expressing constraints on the hyperspace of possible joint probability distributions is used. In this approach, some of the conditional probabilities or (clique) potentials may be elicited from experts or learned from data, where for others, only qualitative information is available.

In this section, we take a similar approach, where we sample the unknown potentials from the factorisation of a given chain graph (cf. Equations (8) and (9)). Without any

further constraints on the hyperspace, the possibility that a positive influence exists between two variables is just as high that a negative influence exists. In fact, when sampling uniformly using a sufficient sample size, the second-order distribution of the value $P(b \mid a, X) - P(b \mid \bar{a}, X)$ resembles a normal distribution with zero mean. The same holds for synergies.

Instead of sampling the full joint probability distributions and then establishing whether the distribution is consistent with the qualitative influences, the potentials can be sampled more efficiently using Lemma 4, as this shows that influences impose *local* constraints on the potentials. An influence can thus be introduced into the hyperspace efficiently by omitting those samples that violate the constraints. Likewise, synergies can be stated in terms of constraints on the local potentials using the following proposition.

Proposition 7. *Given a chain graph* G *containing vertices* A_1, A_2, *and* B, *with* $A_1, A_2 \in \mathrm{bd}(B)$ *and* B *an element of a component* C, *it holds that a positive additive synergy* $Y^+(\{A_1, A_2\}, B)$ *exists if and only if*

$$\phi_C(a_1, a_2, b) + \phi_C(\bar{a}_1, \bar{a}_2, b) \geqslant \phi_C(a_1, \bar{a}_2, b) + \phi_C(\bar{a}_1, a_2, b)$$

and, likewise, a positive product synergy $X^+(\{A_1, A_2\}, b)$ *exists if and only if*

$$\phi_C(a_1, a_2, b) \cdot \phi_C(\bar{a}_1, \bar{a}_2, b) \;\geqslant\; \phi_C(a_1, \bar{a}_2, b) \cdot \phi_C(\bar{a}_1, a_2, b)$$

with

$$\phi_C(A_1, A_2, B) = \frac{\prod_{M \in \mathcal{M}_B} \varphi(A_1, A_2, B)}{\sum_B \prod_{M \in \mathcal{M}_B} \varphi(A_1, A_2, B)}$$

Proof. Similar to Lemma 4, note that:

$$P(B \mid A_1, A_2, \mathrm{fa}(B) - A_1 A_2 B) = \frac{Z^{-1}(\mathrm{pa}(C)) \Big(\prod_{\mathcal{M}_C} \varphi_M(A_1, A_2, B) \Big) P(\mathrm{pa}(C))}{\sum_B Z^{-1}(\mathrm{pa}(C)) \Big(\prod_{\mathcal{M}_C} \varphi_M(A_1, A_2, B) \Big) P(\mathrm{pa}(C))}$$

Here, $\mathrm{pa}(C)$ and $Z^{-1}(\mathrm{pa}(C))$ do not depend on B, so this simplifies to:

$$P(B \mid A_1, A_2, \mathrm{fa}(B) - A_1 A_2 B) = \frac{\prod_{\mathcal{M}_C} \varphi_M(A_1, A_2, B)}{\sum_B \prod_{\mathcal{M}_C} \varphi_M(A_1, A_2, B)}$$

Finally, we can put those potentials that do not depend on B in front of the summation and obtain:

$$P(B \mid A_1, A_2, \mathrm{fa}(B) - A_1 A_2 B) = \frac{\prod_{M \in \mathcal{M}_B} \varphi_M(A_1, A_2, B)}{\sum_B \prod_{M \in \mathcal{M}_B} \varphi_M(A_1, A_2, B)}$$

Call this expression $\varphi_C(A_1, A_2, B)$. Then, the claims readily follow from the definitions of the synergies. □

Procedure 1 *sample-distribution*(potentials: $\varphi_{\mathtt{known}}, \varphi_{\mathtt{unknown}}$; qualitative constraints: C)

satisfied = FALSE
while NOT *satisfied* **do**
  for $\varphi_M \in \varphi_{\mathtt{unknown}}$ **do**
    $\phi_M \leftarrow$ *sample potential randomly* > 0
  end for
  satisfied = TRUE
  for $C_i \in C$ **do**
    satisfied $\leftarrow$ *satisfied* $\wedge$ *check-constraint*$(\varphi_{\mathtt{known}} \cup \{\varphi_M\}, C_i)$
  end for
end while
$\phi_{known} \leftarrow \phi_{known} \cup \{\varphi_M\}$
return φ_{known}

Procedure 2 *check-constraint*(potentials: φ; constraint: $S^{\delta}(A, B)$ *or* $Y^{\delta}(\{A_1, A_2\}, B)$ *or* $X^{\delta}(\{A_1, A_2\}, B)$)

satisfied = FALSE
if *type* = *influence* **then**
  determine all $\varphi_M \in \varphi$ *for which* $M \in M_{AB}$
  satisfied $\leftarrow$ *check* $\{\varphi_M\}$ *with Lemma 4*
end if
if *type* = *synergy* **then**
  determine all $\varphi_M \in \varphi$ *for which* $M \in M_B$
  satisfied $\leftarrow$ *check* $\{\varphi_M\}$ *with Proposition 7*
end if
return *satisfied*

Given these properties, distributions can be sampled (cf. Procedure 1) that satisfy the qualitative constraints (cf. Procedure 2). Then, using these samples, second-order distributions of arbitrary marginal distributions can be derived in a straightforward manner. While typically the marginals range over the whole $[0, 1]$ interval, the qualitative constraints alter the shape (e.g., the mean and variance) of the distribution, which can then be used to draw conclusions from the model.

Using patient data electronic health records of general practices in the Netherlands – the LINH data (see Chapter 3 for details) – we were able to produce realistic qualitative and quantitative information for our running example. Table 8.2 shows the contingency table for familial hypercholesterolaemia, obesity, antidiabetic therapy, high total cholesterol, and high blood glucose levels measured by HbA1c. The prevalences are 0.28%, 1.4%, 7.8%, 4.1%, and 5.0% respectively. This is in line with numbers known for the Dutch population, except for obesity. Research on chronic disease prevalences in the Netherlands [79] showed that the prevalence of obesity is a tenfold of the one found here. The difference might be explained by the fact that in most cases obesity is only diagnosed when the patient specifically asks for a treatment. In our experiment we will

	$\overline{\mathrm{ch}}, \overline{\mathrm{gl}}$	$\overline{\mathrm{ch}}, \mathrm{gl}$	$\mathrm{ch}, \overline{\mathrm{gl}}$	ch, gl	$S(\mathrm{Ch}, \mathrm{Gl})$	$S(\mathrm{Gl}, \mathrm{Ch})$
$\overline{\mathrm{f}}, \overline{\mathrm{ob}}, \overline{\mathrm{th}}$	160849	764	2481	1487	0.645	0.370
$\overline{\mathrm{f}}, \overline{\mathrm{ob}}, \mathrm{th}$	7182	3585	176	2845	0.419	0.609
$\overline{\mathrm{f}}, \mathrm{ob}, \overline{\mathrm{th}}$	1876	32	126	86	0.666	0.389
$\overline{\mathrm{f}}, \mathrm{ob}, \mathrm{th}$	99	146	13	139	0.372	0.319
$\mathrm{f}, \overline{\mathrm{ob}}, \overline{\mathrm{th}}$	381	4	26	23	0.788	0.459
$\mathrm{f}, \overline{\mathrm{ob}}, \mathrm{th}$	4	19	3	15	0.013	0.007
$\mathrm{f}, \mathrm{ob}, \overline{\mathrm{th}}$	24	1	3	3	0.639	0.460
$\mathrm{f}, \mathrm{ob}, \mathrm{th}$	1	1	1	1	0.000	0.000

Table 8.2: Contingency table for familial hypercholesterolaemia (f), obesity (ob), antidiabetic therapy (th), high total cholesterol (ch), and high blood glucose levels (gl) measured by HbA1c. Patient counts are derived from electronic health records of 82 general practices in the Netherlands. The two most right columns show the evaluation of a possible influence between high total cholesterol and high blood glucose levels in the context (Z) of that specific row, i.e., $S(X, Y) = P(x \mid y, Z) - P(x \mid \bar{y}, Z)$.

use the prevalence known from the literature. Table 8.2 also contains the evaluation of a possible influence between high total cholesterol and high blood glucose levels, showing that $P(gl \mid ch, F, Ob, Th) - P(gl \mid \overline{ch}, F, Ob, Th) \geqslant 0$ for any context (instantiation) of F, *Ob* and *Th*. This implies a positive influence between total cholesterol and blood glucose levels, i.e., $S^{+}(Ch, Gl)$. It should be noted that the zero-influences of the last row are probably a coincidence due to the small numbers observed there.

Now, consider the quantitative and qualitative information available in Figure 8.9a, representing the information derived from both the patient data and the literature. We assume that high total cholesterol strongly correlates with lipid disorder (in fact, this is one of the most important measures used to diagnose this disorder). The second-order distribution of high cholesterol (*Ch*) within the general population, based on $10,000$ samples, is shown in Figure 8.9b.

Interventions on *Th* and *Ob* yield the second-order distribution in Figure 8.9d, 8.9d, 8.9e, and 8.9f. In case $Ob = true$ the second-order distributions shift to the right, and in case of $Ob = false$ the second-order distributions shift to the left. The opposite holds for *Th*, suggesting with high confidence that diabetic therapy is also beneficial to reduce cholesterol levels. Note that this has been derived without any quantitative information about the chain component containing *LD* and *DM*.

Table 8.3 compares the second-order distribution of high total cholesterol levels for the different interventions on obesity and anti-diabetic therapy. One can see that anti-diabetic therapy has a small benefit to reduce cholesterol levels for both obese and non-obese patients. However, the differences between obese and non-obese people are much larger, suggesting that an additional reduction of weight in combination with diabetic therapy is even more beneficial to reduce cholesterol levels.

These conclusions cannot be derived from the data directly. Since the data used here is derived from general practices, untreated diabetics are sparse, which makes the negative influence from *Th* to *DM* impossible to detect. In fact, in the data, the presence

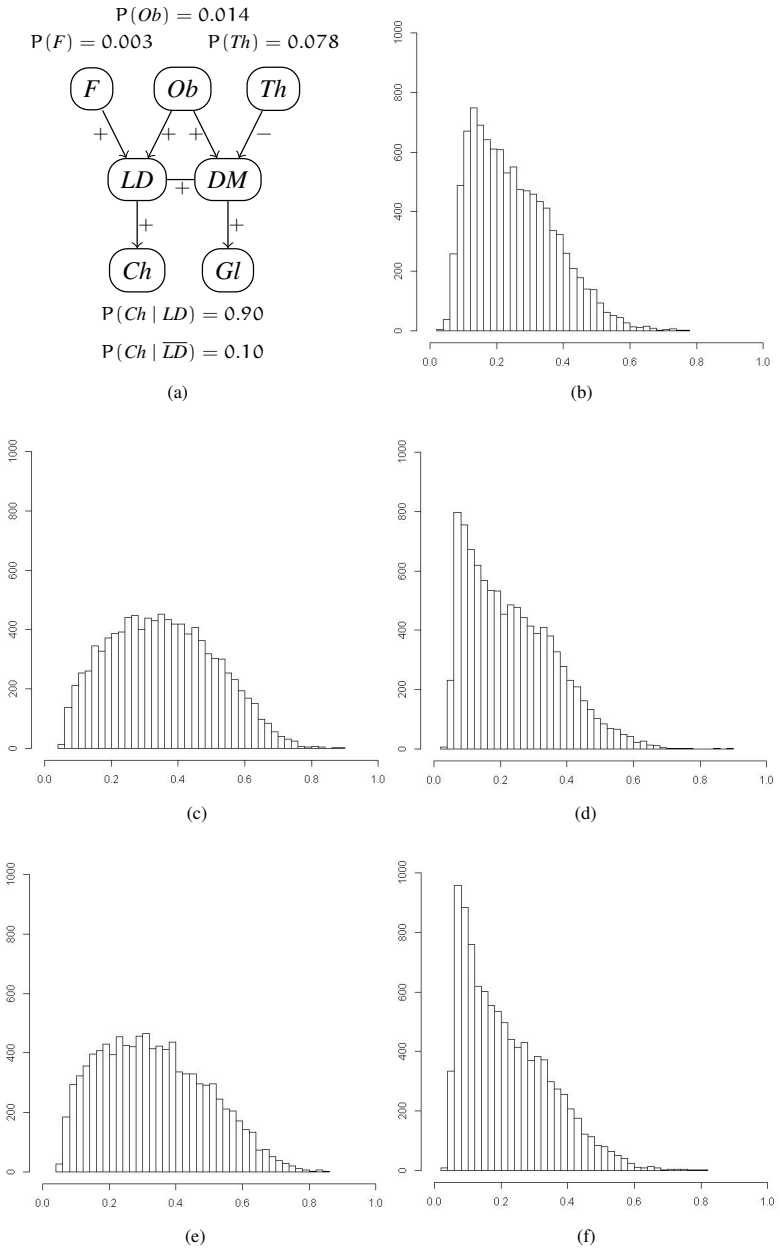


Figure 8.9: Qualitative and quantitative information (a) of Figure 8.5(i), and second-order probability distributions of Ch in general (b), and in the presence of specific interventions, i.e., $Ob, \overline{Th}$ (c), $\overline{Ob}, \overline{Th}$ (d), Ob, Th (e), and $\overline{Ob}, Th$ (f).

	Probabilities being compared		Percentage
	P_1	P_2	$P_1 < P_2$
Effect of therapy	$P(Ch \mid Ob, Th)^{e}$	$P(Ch \mid Ob, \overline{Th})^{c}$	62.72
	$P(Ch \mid \overline{Ob}, Th)^{f}$	$P(Ch \mid \overline{Ob}, \overline{Th})^{d}$	64.24
Effect of obesity	$P(Ch \mid \overline{Ob}, \overline{Th})^{d}$	$P(Ch \mid Ob, \overline{Th})^{c}$	85.22
	$P(Ch \mid \overline{Ob}, Th)^{f}$	$P(Ch \mid Ob, Th)^{e}$	84.59
Combined effect	$P(Ch \mid \overline{Ob}, Th)^{f}$	$P(Ch \mid Ob, \overline{Th})^{c}$	87.81

Table 8.3: Comparison of second-order distribution of high total cholesterol for different interventions on obesity and antidiabetic therapy. The superscripts c, d, e, and f, refer to the second-order distributions in Figure 8.9.

of antidiabetic therapy makes higher blood glucose levels more likely. However, from the literature it is known that diabetic therapy *reduces* glucose levels. We argue that it is better to incorporate this as qualitative information into the model rather than learning it from the data, thereby avoiding the biased relation between *Th* and *DM*.

8.7 CONCLUSIONS

The work described in this chapter started off with the wish to exploit qualitative information in modelling dynamic systems in a state of equilibrium. This brought us to the development of an extension of qualitative probabilistic networks (QPNs) towards qualitative chain graphs (QCGs). We were able to obtain generalised definitions of qualitative influences and additive synergies. Product synergies still express intercausal reasoning in QCGs, although some of the intercausal reasoning in chain graphs cannot be captured by product synergies. Furthermore, we studied sign propagation for singly-connected chain graphs; similar ideas can be explored for multiply-connected networks. From the point of view of sign propagation, the key difference between QPNs and QCGs are the trails over which signs are propagated. This makes it clear that it matters whether qualitative relationships between variables in a probabilistic network are defined with respect to a line or an arc as this changes the dependences between variables. Elicitation of qualitative relationship should, therefore, go hand in hand with graph structure elicitation.

The value of QPNs for modelling biomedical problems has been recognised before [138, 117]. Although qualitative reasoning with QCGs has similar limitations as with QPNs, we showed that by exploiting qualitative constraints on the chain graph potentials we are able to estimate arbitrary influences and synergies in the chain graph, i.e., by means of the second-order distribution of the marginal probabilities. For example, we can provide the second-order distribution of $P(b \mid a, X) - P(b \mid \bar{a}, X)$, and decide up to a certain significance level how likely it is that an influence $S^{+}(A, B)$ exists. Therefore, we believe that this result may have a real practical bearing on areas

such medicine: without knowing the exact joint probability distribution, we are still able to draw qualitative conclusions on the dynamics that exist within a model.

One apparent limitation of this work is that we restrict the theory to dealing with binary variables. The qualitative properties of QPNs are based on the concept of first-order dominance [221], which does capture ordinal relationships between probability distributions, in particular that a positive influence makes the higher values of that variable more likely, i.e., $P(c \geqslant c_0 \mid a_1, x) \geqslant P(c \geqslant c_0 \mid a_2, x)$ for all $a_1 > a_2$, c_0, and context x. It is not difficult to see that all our results generalise to arbitrary discrete distributions, because we can always encode a discrete distribution using binary variables. In a naive way, a random variable A with n values may be represented by binary variables for each $a_i > a_j$, which is true if a_i holds and false if a_j holds. Subsequently, factors can be added to the network to make sure that exactly one a_i is true. For continuous random variables, the situation is more difficult. Most of the proofs are similar by just taking $A = \textit{true}$ as shorthand for a_1 and $A = \textit{false}$ as shorthand for a_2, such that $a_1 > a_2$. However, for proofs that depend on the factorisation, such as Lemma 4, it is required that the distribution is faithful to the graph. While there exists research for the discrete and Gaussian case [153, 154], other distributions have not been studied thoroughly.

From a theoretical point of view there are some clear directions for further research. For QCGs, sign propagation can be investigated for general chain graphs. Another interesting line of research is to investigate qualitative abstractions of other probabilistic graphical models, such as acyclic directed mixed graphs, which have directed and bi-directed edges. The latter can be used to represent hidden common causes (see [169] for its Markov properties). To improve inference in the sampling algorithm, the sampling of potentials may be improved by exploiting Monte Carlo methods which take into account bounds on the hyperspace (e.g. based on [191]).

With respect to the medical problems we have used for illustrative purposes, we aim to apply this formalism in a study on diabetes and cardiovascular comorbidities involving multiple feedback systems. Since most physiological systems cannot be measured directly, and relevant parameters are, therefore, mostly absent in large epidemiological datasets that are available, we believe that the methods presented here form a good foundation for developing multiple disease models.

9

DISCUSSION

The goal of the research underlying this thesis was to develop new techniques to analyse interactions between different entities of all kinds based on data. The techniques we developed used ideas from the area of probabilistic graphical models; healthcare was the field used to validate the novel techniques. Sometimes, important special features of healthcare data guided the development of the techniques. The entities and their interactions on which most of the research focused were diseases, and, in particular the presence of multiple diseases in patients, also called multimorbidity. The value of having available probabilistic models of multiple diseases is that these can act as a foundation for clinical guidelines of multimorbidity that can be tailored to the requirements of individual patients.

Firstly, we will summarise and discuss the individual scientific achievements of the Chapters 3 through 8, and place their contribution in a broader context. Secondly, we provide some directions for future research, both for probabilistic graphical models and multimorbidity.

9.1 MAIN CONTRIBUTIONS

9.1.1 *Bayesian networks as means to capture interactions*

At the beginning of the research we asked ourselves the question how interactions between entities, in particular when they are stochastic in nature, can be best described. In most of the empirical scientific literature, the interaction between two variables is described by the odds ratio, whereas interactions involving more variables are described using regression equations. Furthermore, different complementary measures are used to obtain the full picture of the presence of interactions. Usually it is hard to fit these together to obtain a single, comprehensive view. Often an interaction between two variables can be clarified in terms of a third variable; more complex interactions also exist. These interactions can be seen as conditional dependences and independences involving multiple variables, which can be expressed quite naturally by Bayesian networks. A further advantage offered by Bayesian networks is that dependences and independences can be hand crafted, learned from data – Bayesian-network structure learning –, or one can combine manual modelling and learning. Throughout this thesis, we have explored the usefulness of Bayesian networks in modelling disease interactions.

Bayesian-network learning was explored in several chapters: in Chapter 4 to better understand the mutual interaction between chronic diseases, in Chapter 5 to obtain in-

sight into the behaviour of different Bayesian-network structure-learning algorithms on large datasets, and in Chapters 6 and 7, to detect static and temporal interactions between disorders. A large dataset with patient data from general practices in the Netherlands was used for that purpose. In comparison to other research in this area that focuses on pairwise associations only, more detailed insights in the interaction, temporal and atemporal, of chronic diseases were obtained. The main achievement of this research is that Bayesian-network structure learning methods can significantly contribute to unravelling the intricate interactions that are hidden in clinical data. Also of interest is that available domain knowledge can be exploited in the learning process to guide model construction, also known as supervised structure learning.

9.1.2 *Bayesian networks for multilevel analysis*

Large datasets often come from multiple sources, thereby introducing a certain bias in the data. In these cases, multilevel regression is often used for analysis. We introduced a new Bayesian-network framework in Chapter 6 that combines probabilistic graphical modelling with multilevel analysis, called multilevel Bayesian networks (MBNs). Using both synthetic and real-world data we showed that the results obtained by MBNs match those of multilevel regression equations in predictive power with the additional advantages offered by Bayesian networks. In Chapter 7 we extended the new framework with a temporal dimension, to be able to analyse longitudinal observational data, giving rise to multilevel temporal Bayesian networks (MTBNs). They were used to model longitudinal change in multimorbidity and showed their usefulness in determining the temporal effects of interactions.

The hierarchical models described in the Chapters 6 and 7 are generalisable to any observational longitudinal dataset containing a hierarchical structure. In that respect, they are suitable for application in population studies of any kind. Returning to the Arctic example from the introduction; the Polar bear population is divided into four major regions, i.e., divergent ice, convergent ice, seasonal ice, and archipelago regions. In turn, these major regions are divided into nineteen minor regional subpopulations. In the work of Amstrup [4] a Bayesian network model was developed for each major region. One can notice that the data consists of variables acting on a regional level, e.g., subpopulation counts, and variables acting on a global level, e.g., worldwide greenhouse gas concentrations. Together with the temporal aspects – yearly and seasonally counts – of the data, the models used in [4] could be merged into an MTBN, tightening the regional effects together by global effects, which would make predictions more accurate.

9.1.3 *Chain graphs as means to model feedback systems*

In Chapter 8 we took a different approach to model interactions: we adopted a *qualitative* viewpoint to understand interactions present in data. Qualitative abstractions of Bayesian networks are known as qualitative probabilistic networks (QPNs); interactions between variables are expressed by signs. In many real-life systems, interactions often participate in feedback loops. Unfortunately, feedback loops cannot be han-

dled by a Bayesian network. However, hybrid probabilistic graphical models, known as chain graphs, can be used to model feedback loops. In the thesis, we adopted QPN theory to develop the new concept of qualitative chain graphs (QCGs) and demonstrated their usefulness in the context of medicine.

Although qualitative reasoning with QCGs has similar limitations as for QPNs, we showed that by exploiting qualitative constraints on the chain graph potentials, we are able to estimate arbitrary influences and synergies. Therefore, we believe that this result may have a real practical bearing on areas such as medicine: without knowing the exact joint probability distribution, we are still able to draw qualitative conclusions on the dynamics that exist in a system. However, the application of QCGs is not restricted solely to the medical field. Many natural processes, such as in chemistry or economics, are characterised by feedback loops that maintain an equilibrium. QCGs may be applicable to these domains as well.

9.1.4 *Multimorbidity and disease interactions*

In Chapters 3 and 4 we especially explored the question whether the many different, and often informal and imprecise concepts used in multimorbidity research as found in the clinical and epidemiological literature can be translated into the formal framework of Bayesian networks. The probabilistic represenations of multimorbidity allows making a distinction between: (1) aetiological models of multimorbidity, and (2) models for reasoning about the clinical impact of multimorbidity. The latter models can be further divided into diagnostic, prognostic, and therapeutic interaction models. The advantage of expressing multimorbidity concepts into probabilistic form is that we are in this way able to model dependences between multiple disease variables. Such models are crucial to adequately manage the illness of patients with multimorbidity.

Speaking of aetiological models, in real-life, some people have to deal with cancer multiple times in their lifetime. This is either due to metastatis, common risk factors, such as smoking, or just bad luck. For such patients, we explored the possible interactions between pairwise observed cancers, using a novel measure called *critical factors* in Chapter 5. It is defined as the minimal set of risk factors in a disease network, that is needed to explain an *indirect* interaction, and the results for the oncological data demonstrate that the measure is a useful concept. Moreover, being defined as a property of a Bayesian network, the use of critical factors is not solely limited to the field of oncology.

We used multilevel Bayesian networks, both static and dynamic, for the analysis of disease interaction and clinically meaningful results were obtained in this way (Chapters 6 and 7). The developed methodology is not restricted to the particular set of diseases studied in this thesis. In fact, since the method takes the epidemiologic bias introduced by obtaining patient data from multiple sources into account, the effect of any combination of diseases on any set of other diseases can be evaluated. Therefore, we recommend the usage of multilevel temporal Bayesian networks to further explore multimorbidity statistics in a broader sense. For example, analysis of the musculoskeletal disorders revealed multiple interactions with chronic diseases of many other organic systems. It would be valuable to describe which of the interactions are the most responsible for this phenomenon.

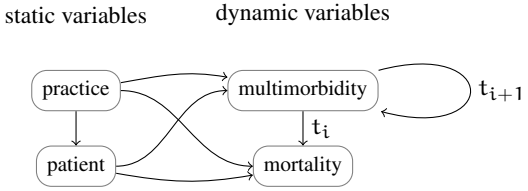


Figure 9.1: Schema of a multimorbidity–mortality model.

9.2 FUTURE RESEARCH

9.2.1 *Probabilistic graphical models*

Although there has been a lot of research in Bayesian network structure learning, very few of the papers on this subject deal with the problem of learning from very large real-world datasets, that is, including both many records and many variables. Little is known, therefore, about which algorithms perform best under these circumstances. Our experience with the the LINH dataset was that search-and-score-based algorithms with the AIC as scoring method gave the best results. In preliminary experiments, we were unable to significantly improve the results using hybrid search-and-score-based with constraint-based algorithms, i.e., by restricting the search space using results from a constraint based algorithm. Therefore, we believe that there is a need for further improvement of structure-learning algorithms to deal with large real-world medical datasets.

Multilevel Bayesian networks (MBNs) as proposed in Chapters 6 and 7 offer room for more research with regard to methods for parameter estimation. In this thesis we used MCMC methods, but there might be more efficient ways to estimate parameters.

In Chapter 8 we developed the initial foundations of qualitative chain graphs (QCGs) as a means to model feedback systems in a qualitative fashion. There are many directions for further research here, such as further investigation of sign propagation in QCGs, or improving the inference that was based on second order distributions. Furthermore, one could develop qualitative abstractions of other probabilistic graphical models, such as maximal ancestral graphs, which have undirected, directed and bidirected edges. It would also be interesting to use the formalism to study multiple feedback systems involved in multiple chronic diseases.

9.2.2 *Multimorbidity and mortality*

In the medical research described in the thesis, we have not investigated the relationship between multimorbidity and mortality, although this is clearly a relevant research topic. The MTBN formalism developed in Chapter 7 can be easily extended with a vertex in each timeslice representing the mortality rate for that particular timeslice. Figure 9.1 shows a schematic picture of an MBN that includes mortality rates.

In preliminary research, we have applied this schema to data from the LINH database, as it contains information about mortality and residential movements. Although not

Urbanisation level	Follow-up 3 years	Follow-up 5 years
very high	0.48 (0.2)	1.9 (0.4)
high	0.50 (0.2)	2.2 (0.5)
moderate	0.50 (0.2)	2.1 (0.5)
low	0.42 (0.2)	1.8 (0.4)
rural	0.49 (0.2)	2.1 (0.5)

(a) Urbanity

Age Group	Follow-up 3 years	Follow-up 5 years
35-50 years	0.34 (0.2)	1.7 (0.4)
50-65 years	0.48 (0.2)	2.0 (0.4)
65-80 years	0.72 (0.3)	2.6 (0.5)
80+ years	0.74 (0.3)	3.0 (0.5)
Overall	0.48 (0.2)	2.0 (0.4)

(b) Age

Table 9.1: Demographic differentiations of mortality rates (in % with standard errors).

entirely accurate in its completeness – something which can be checked against the national registry of births and deaths – the numbers can shed light on the relation between multimorbidity and mortality. Tables 9.1 and 9.2 show that there is enough differentiation in the dataset concerning mortality rates with respect demographics and clusters of chronic diseases.

If we divide the LINH data into several timeslices for the clusters distinguished in Table 9.2, and apply structure learning – guided by the high-level structure of Figure 9.1 – we obtain a preliminary view on how multimorbidity is related to mortality. Figure 9.2 shows that in particular cardiovascular, endocrine, and musculoskeletal disorders are associated with disorders of all kinds in the next time slice. This is in line with the results described in Chapter 3. In contrast, the disease clusters associated with mortality are the respiratory, psychiatric, oncologic, and cardiovascular disease clusters.

These results suggest that multimorbidity affects both quality of life and mortality. However, not necessarily the same diseases are involved in these two processes. After all, prolonging the *length* of life by medical care is not the same as prolonging the *quality* of life by medical care, and the relation between multimorbidity, mortality, and quality of life requires further investigation.

9.3 FINAL NOTE

In this thesis, we have provided new methods for probabilistic graphical models that can be used in the analysis of disease interactions. This was done by extending and combining particular concepts – both quantitative and qualitative – from existing probabilistic graphical models. We applied them to the medical area of multimorbidity, but the ideas are certainly generalisable to other areas of scientific research. As an example we briefly discussed a simplified model coming from the environmental modelling field, i.e., a model of the Arctic summer sea-ice decline.

Besides, we provided some preliminary ideas on how to use network models in building clinical guidelines, keeping in mind that a patient can face multiple diseases at the same time. However, their actual use in practice has still to be explored. We are convinced that network models can be very useful to link together the statistics, on

Disease Cluster	Follow-up 3 years	Follow-up 5 years
Malignity	3.5 (0.6)	9.9 (0.9)
Cardiovascular	1.0 (0.3)	3.7 (0.6)
Respiratory	0.9 (0.3)	3.6 (0.6)
Psychiatric	0.9 (0.3)	3.3 (0.6)
Endocrine	0.7 (0.3)	2.7 (0.5)
Ophthalmologic	0.7 (0.3)	2.7 (0.5)
Otolaryngologic	0.6 (0.2)	2.6 (0.5)
Urogenital	0.6 (0.2)	2.6 (0.5)
Haematologic	0.5 (0.2)	2.6 (0.5)
Musculoskeletal	0.5 (0.2)	2.4 (0.5)
Gastroenterologic	0.5 (0.2)	2.3 (0.5)
Dermatologic	0.5 (0.2)	2.3 (0.5)
Neurologic	0.5 (0.2)	2.3 (0.5)
Congenital	0.4 (0.2)	2.0 (0.4)
Tuberculosis/HIV	0.3 (0.2)	1.8 (0.4)

(a) Clusters

Cluster Combination	Follow-up 3 years	Follow-up 5 years
Malignity + Cardiovascular	5.2 (0.7)	12.5 (1.0)
Malignity + Respiratory	2.2 (0.7)	8.1 (1.1)
Malignity + Psychiatric	3.4 (0.7)	9.6 (1.1)
Malignity + Endocrine	3.2 (0.6)	9.9 (1.0)
Malignity + Musculoskeletal	3.4 (0.6)	11.5 (0.9)
Cardiovascular + Respiratory	1.9 (0.4)	5.9 (0.7)
Cardiovascular + Psychiatric	1.9 (0.4)	5.7 (0.7)
Cardiovascular + Endocrine	1.0 (0.3)	3.8 (0.6)
Cardiovascular + Musculoskeletal	1.0 (0.3)	3.8 (0.6)
Respiratory + Psychiatric	1.0 (0.4)	3.9 (0.7)
Respiratory + Endocrine	1.0 (0.4)	4.8 (0.6)
Respiratory + Musculoskeletal	1.5 (0.3)	4.6 (0.6)
Psychiatric + Endocrine	1.3 (0.3)	4.3 (0.6)
Psychiatric + Musculoskeletal	1.2 (0.3)	4.3 (0.6)
Endocrine + Musculoskeletal	0.7 (0.3)	2.7 (0.5)

(b) Combinations

Table 9.2: Mortality rates for disease clusters and combinations (in % with standard errors).

which individual clinical guidelines are based, with the clinical reasoning. This would create a sound foundation for the development of clinical guidelines that are more suitable for the multimorbidity patient.

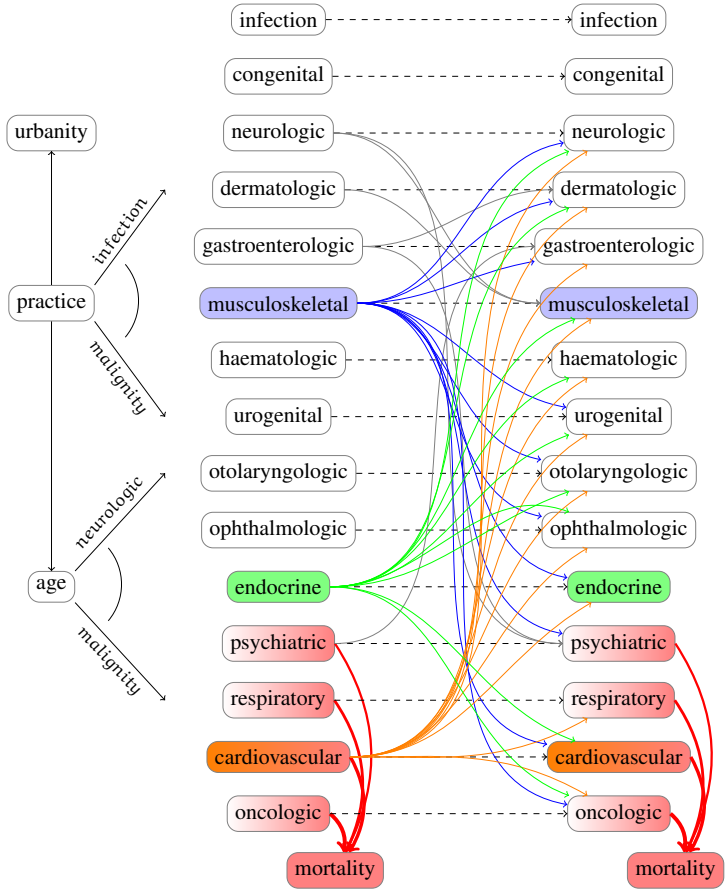


Figure 9.2: Preliminary multimorbidity–mortality model for disease clusters.

A

APPENDIX

ADDITIONAL MATERIALS AND RESULTS ON MULTILEVEL TEMPORAL BAYESIAN NETWORKS

Here we provide extra figures and tables belonging to the analysis that was made in Chapter 7. Figure A.1 shows the global structure of the MTBN used. A distinction is made between cardiovascular health risk determinants, chronic diseases in relation to practice, and patient characteristics. Besides incidence rates per time slice, the probability distribution of the MTBN allows us to detect interactions between disorders in the context of health risks as well.

Model 3 shows the pseudocode for parameter learning of the MTBN in WinBUGS. The number of patients is denoted by N, the number of practices by K. The hyper parameter γ is mostly set to zero initially, but one can also choose other values. The vectors of mean Γ is also typically set to zero. The variance σ is mostly drawn from a gamma distribution with parameters 0.001 and 0.001, and likewise the co-variance matrix Σ is initialised. Here, the local probability distributions are estimated using a logistic link function for each variable (the variables D_t^{i*} are just intermediate variables to accomplish this). For readability, the health risk factors were modelled in the pseudocode as they were diseases as well (i.e., for some D_t^i it holds that $D_t^i = R_t^k$). The data to be provided along with the model contains the raw data on urbanity (Ur), age Age, gender Gen, and diseases (D_t^i). In the model P and Q stand for any parent of a disease D_t^i not being a higher level variable (urbanity), i.e., $P, Q \in parents(D_t^i) \setminus \{Ur\}$. Note that age and gender can also be a parent of D_t^i. The double sums in the logit equation model the interaction terms between parents on a disease.

The results from the model in Chapter 7 are aggregated into more comprehensive figures in the chapter itself. Here we show the numbers that were used to build these figures. Table A.1 shows which diseases were used in the final model. Table A.2 shows the baseline characteristics with respect to the urbanization level of the practice location, and the age and gender of the patient. Figure A.2 shows the final MTBN obtained by structure learning. All variables depend on age and gender; for clarity these dependencies have been omitted. Table A.3 and A.4 show prevalences differentiated for age and urbanity. The baseline and 5-year follow-up results differentiated for urbanity are also visualised in Figure 7.1 of the Chapter 7. Table A.5 and A.6 of this supplementary show prevalences of comorbid patterns, and they are aggregated into Table 7.2 of 7. Finally, Figure A.3 of this supplementary shows detailed information for diabetics.

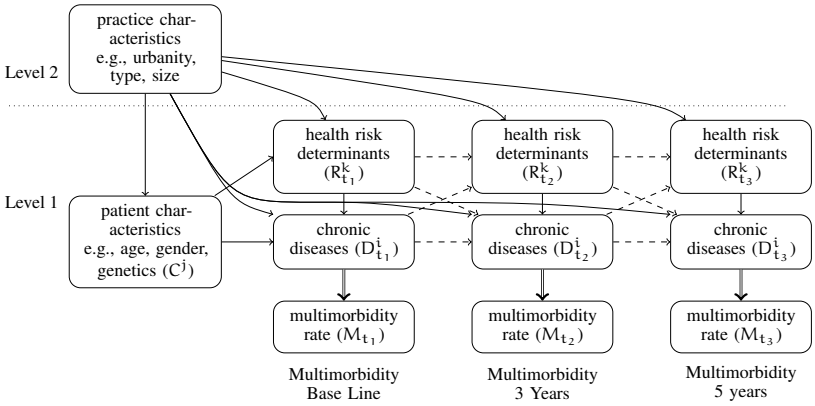


Figure A.1: High level structure of the multimorbidity model. Solid arcs represent associations between diseases within one time slice. Dashed arcs represent associations between diseases for different time slices. A double arc represents a deterministic relation. 4) The dotted line separates the practice variables from the patient variables.

Model 3 *pseudocode WinBUGS*

```
model;
{
  for (n in 1:N)
  {
    # For all values of i and t:
    D^i_t[n] ~ dbern(D^i*_t[n])
    # Variables without practice variation:
    logit(D^i*_t[n]) ← γ^i_t0 + Σ_{P∈parents} γ^i_tp P[n] +
                       Σ Σ_{P,Q∈parents} γ^i_tpq P[n]Q[n]
    # Variables with practice variation:
    logit(D^i*_t[n]) ← γ^i_t0[Ur[n]] + Σ_{P∈parents} γ^i_tp[Ur[n]]P[n] +
                       Σ Σ_{P,Q∈parents} γ^i_tpq[Ur[n]]P[n]Q[n]
  }
  # Priors non-practice related parameters
  γ^i_t0 ~ dnorm(γ, σ)
  γ^i_tp ~ dnorm(γ, σ)
  γ^i_tpq ~ dnorm(γ, σ)
  # Priors practice related parameters
  γ^i_t0[1 : K] ~ dmnorm(Γ, Σ)
  γ^i_tp[1 : K] ~ dmnorm(Γ, Σ)
  γ^i_tpq[1 : K] ~ dmnorm(Γ, Σ)
}
```

Disease	ICPC codes	Description
Obesity	T82, T83	Obesity and overweight
Dyslipidemia	T93	Lipid disorders
Hypertension	K86, K87	Complicated and un-complicated hypertension
Diabetes Mellitus	T89, T90	Insulin and non-insulin dependent diabetes mellitus
Ischemic Heart Disease	K74, K75, K76	Angina pectoris, coronary atherosclerosis, myocardial infarct
Heart Failure	K77	Heart failure
Stroke	K89, K90	Transient cerebral ischaemia and cerebrovascular accident
Nephropathy	U88, U99	Glomerulonefritis/nephrosis, renal failure
Retinopathy	F83	Retinopathy
Other Diseases		
– Malignant	A79, B72, B73, D74, D75, D76, D77, L71, N74, R84, R85, S77, T71, U75, U76, U77, W72, X75, X76, X77, Y77, Y78	
– Non-Malignant	A70, B90, D85, D86, D94, F84, F92, F93, F94, H84, H85, H85, K73, L83, L84, L85, L86, L88, L89, L90, L91, L95, N85, N86, N87, P15, P70, P72, P74, P76, P79, P85, T06, R91, R95, R96, S87, S88	

Table A.1: ICPC codes used to identify the diseases that were used in the model.

Urbanization	Age Group				Gender		
level	35-49 yr	50-64 yr	65-79 yr	80+ yr	Male	Female	Total
very high	16580	10631	5957	2369	17446	18091	35537
high	16925	13297	7195	2425	18888	20956	39844
moderate	13675	9738	4709	1167	14231	15058	29289
low	16787	14147	7438	1735	19686	20421	40107
rural	15869	13047	6766	1937	18799	18820	37619
Total	79836	60860	32067	9633	89050	93346	182396

Table A.2: Age and gender of the study population at baseline by urbanization level of the practice location. Urbanization varies from 'very low' (less than 500 addresses per km^2), also denoted as 'rural', to 'very high' (more than 2500 addresses per km^2).

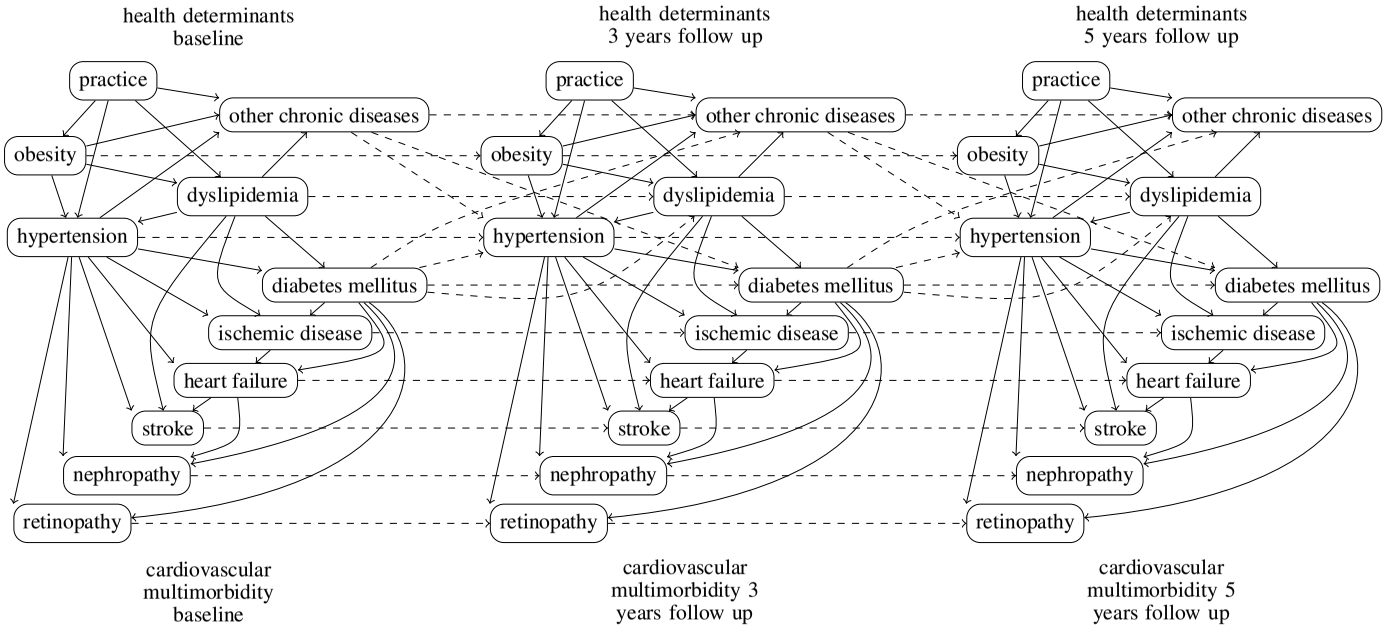


Figure A.2: Multilevel Bayesian network of health-risk determinants and chronic cardiovascular diseases for persons of 35 years and older, at baseline, after 3 follow-up, and 5 years follow-up. The solid lines represent the associations between diseases within one time slice, dashed lines represent the temporal relation between two different diseases, and dotted lines represent the temporal relation for one disease. For clarity, age, gender, and the overall multimorbidity rates are not shown (see also statistical analysis section).

		Health Risks			Chronic Cardiovascular Diseases						
	Age Group	Obesity	Dyslipi-demia	Hyper-tension	Other Diseases	Diabetes Mellitus	Ischemic Heart D.	Heart Failure	Stroke	Nephro-pathy	Retino-pathy
base	35-50 yr	1.1 (0.3)	1.1 (0.3)	2.9 (0.5)	15.1 (1.1)	1.0 (0.3)	0.3 (0.2)	0.1 (0.1)	0.2 (0.1)	0.1 (0.1)	<.1 (0.1)
line	50-65 yr	1.4 (0.4)	3.8 (0.6)	11.2 (1.0)	20.4 (1.3)	4.4 (0.7)	2.1 (0.5)	0.3 (0.2)	0.7 (0.3)	0.3 (0.2)	<.1 (0.1)
	65-80 yr	1.3 (0.4)	6.3 (0.8)	21.5 (1.3)	28.7 (1.4)	9.7 (0.9)	5.4 (0.7)	1.5 (0.4)	2.3 (0.5)	0.8 (0.3)	0.1 (0.1)
	80+ yr	0.7 (0.3)	3.1 (0.5)	23.8 (1.3)	34.3 (1.5)	10.3 (1.0)	7.6 (0.8)	5.6 (0.7)	4.1 (0.6)	1.4 (0.4)	0.2 (0.1)
	overall	1.2 (0.3)	3.0 (0.5)	10.1 (0.9)	20.3 (1.3)	4.2 (0.6)	2.2 (0.5)	0.7 (0.3)	0.9 (0.3)	0.4 (0.2)	<.1 (0.1)
3 yrs	35-50 yr	2.6 (0.5)	2.8 (0.5)	6.6 (0.8)	32.7 (1.6)	2.5 (2.5)	1.2 (0.3)	0.1 (0.1)	0.5 (0.2)	0.4 (0.2)	0.1 (0.1)
	50-65 yr	3.2 (0.6)	8.7 (0.9)	22.4 (1.3)	42.4 (1.5)	9.2 (0.9)	5.0 (0.7)	0.8 (0.3)	2.0 (0.4)	1.1 (0.3)	0.2 (0.1)
	65-80 yr	2.4 (0.5)	13.2 (1.1)	38.9 (1.5)	56.7 (1.6)	18.0 (1.2)	11.9 (1.0)	4.4 (0.6)	6.0 (0.8)	3.5 (0.6)	0.5 (0.2)
	80+ yr	1.2 (0.3)	7.2 (0.8)	41.9 (1.6)	67.5 (1.5)	19.5 (1.3)	15.7 (1.1)	15.6 (1.1)	11.1 (1.0)	6.7 (0.8)	0.5 (0.2)
	overall	2.7 (0.5)	6.8 (0.8)	19.5 (1.3)	42.0 (1.6)	8.4 (0.9)	5.0 (0.7)	1.9 (0.4)	2.5 (0.5)	1.6 (0.4)	0.2 (0.1)
5 yrs	35-50 yr	3.5 (0.6)	3.7 (0.6)	8.7 (0.9)	39.9 (1.5)	3.2 (0.6)	1.4 (0.4)	0.2 (0.1)	0.7 (0.2)	0.6 (0.2)	0.1 (0.1)
	50-65 yr	4.4 (0.6)	10.8 (0.9)	27.3 (1.4)	51.5 (1.6)	11.1 (1.0)	6.3 (0.8)	1.1 (0.3)	2.8 (0.5)	1.7 (0.4)	0.4 (0.2)
	65-80 yr	3.0 (0.5)	15.0 (1.1)	44.6 (1.6)	66.2 (1.5)	20.9 (1.3)	14.4 (1.1)	6.3 (0.8)	8.0 (0.9)	5.7 (0.7)	0.8 (0.3)
	80+ yr	1.3 (0.4)	7.9 (0.8)	46.0 (1.6)	74.7 (1.4)	21.0 (1.3)	18.9 (1.2)	20.5 (1.3)	14.3 (1.1)	9.2 (0.9)	0.7 (0.3)
	overall	3.6 (0.6)	8.4 (0.9)	23.1 (1.3)	50.3 (1.6)	9.9 (0.9)	6.3 (0.8)	2.7 (0.5)	3.4 (0.6)	2.3 (0.5)	0.4 (0.2)

Table A.3: Base Line Prevalences and Estimated Follow-up Probabilities by Age in % (with standard errors) of Health Risks and Chronic Cardiovascular Multimorbidity.

		Health Risks			Chronic Cardiovascular Diseases						
	Urbanity	Obesity	Dyslipi-demia	Hyper-tension	Other Diseases	Diabetes Mellitus	Ischemic Heart D.	Heart Failure	Stroke	Nephro-pathy	Retino-pathy
base	very high	1.1 (0.3)	1.7 (0.4)	6.6 (0.8)	14.6 (1.1)	3.5 (0.6)	1.9 (0.4)	0.6 (0.2)	0.8 (0.3)	0.3 (0.2)	<.1 (0.1)
line	high	1.7 (0.4)	4.4 (0.7)	12.0 (1.0)	25.0 (1.4)	4.7 (0.7)	2.5 (0.5)	0.8 (0.3)	1.1 (0.3)	0.4 (0.2)	<.1 (0.1)
	moderate	1.4 (0.4)	4.3 (0.6)	12.0 (1.0)	23.1 (1.3)	4.4 (0.6)	2.4 (0.6)	0.7 (0.3)	1.0 (0.3)	0.4 (0.2)	<.1 (0.1)
	low	0.6 (0.3)	2.3 (0.5)	7.9 (0.8)	15.8 (1.1)	3.8 (0.6)	2.0 (0.4)	0.6 (0.2)	0.8 (0.3)	0.3 (0.2)	<.1 (0.1)
	rural	1.4 (0.4)	2.7 (0.5)	12.0 (1.0)	24.4 (1.4)	4.3 (0.6)	2.3 (0.5)	0.7 (0.3)	1.0 (0.3)	0.4 (0.2)	<.1 (0.1)
3 yrs	very high	2.6 (0.5)	5.5 (0.7)	16.6 (1.2)	37.9 (1.5)	7.8 (0.8)	4.6 (0.7)	2.0 (0.4)	2.4 (0.5)	1.5 (0.3)	0.2 (0.1)
	strong	3.1 (0.6)	8.2 (0.9)	21.5 (1.3)	45.9 (1.6)	9.0 (0.9)	5.6 (0.7)	2.1 (0.5)	2.9 (0.5)	1.6 (0.3)	0.2 (0.1)
	moderate	2.8 (0.6)	7.9 (0.9)	20.6 (1.3)	43.6 (1.6)	8.3 (0.9)	5.0 (0.7)	1.8 (0.4)	2.5 (0.5)	1.5 (0.3)	0.2 (0.1)
	low	2.2 (0.5)	6.3 (0.8)	18.0 (1.2)	38.0 (1.5)	8.2 (0.9)	4.8 (0.7)	1.8 (0.4)	2.3 (0.5)	1.5 (0.3)	0.2 (0.1)
	rural	2.9 (0.5)	6.6 (0.8)	21.3 (1.3)	45.1 (1.6)	8.6 (0.9)	5.2 (0.7)	2.0 (1.4)	2.6 (0.4)	1.5 (0.3)	0.2 (0.1)
5 yrs	very high	3.4 (0.6)	7.0 (0.8)	20.0 (1.3)	46.7 (1.6)	9.3 (0.9)	5.8 (0.7)	2.7 (0.5)	3.3 (0.6)	2.2 (0.5)	0.3 (0.2)
	high	4.0 (0.6)	9.8 (0.9)	25.0 (1.4)	53.5 (1.6)	10.5 (1.0)	6.8 (0.8)	2.8 (0.5)	3.7 (0.6)	2.5 (0.5)	0.4 (0.2)
	moderate	3.8 (0.6)	9.3 (0.9)	24.1 (1.4)	51.4 (1.6)	9.8 (0.9)	6.2 (0.8)	2.4 (0.5)	3.3 (0.6)	2.3 (0.5)	0.4 (0.2)
	low	3.1 (0.5)	7.8 (0.8)	21.6 (1.3)	46.9 (1.6)	9.6 (0.9)	6.0 (0.7)	2.5 (0.5)	3.2 (0.6)	2.2 (0.5)	0.3 (0.2)
	rural	3.8 (0.6)	8.1 (0.9)	24.7 (1.4)	52.3 (1.6)	10.1 (1.0)	6.3 (0.8)	2.7 (0.5)	3.5 (0.6)	2.4 (0.5)	0.4 (0.2)

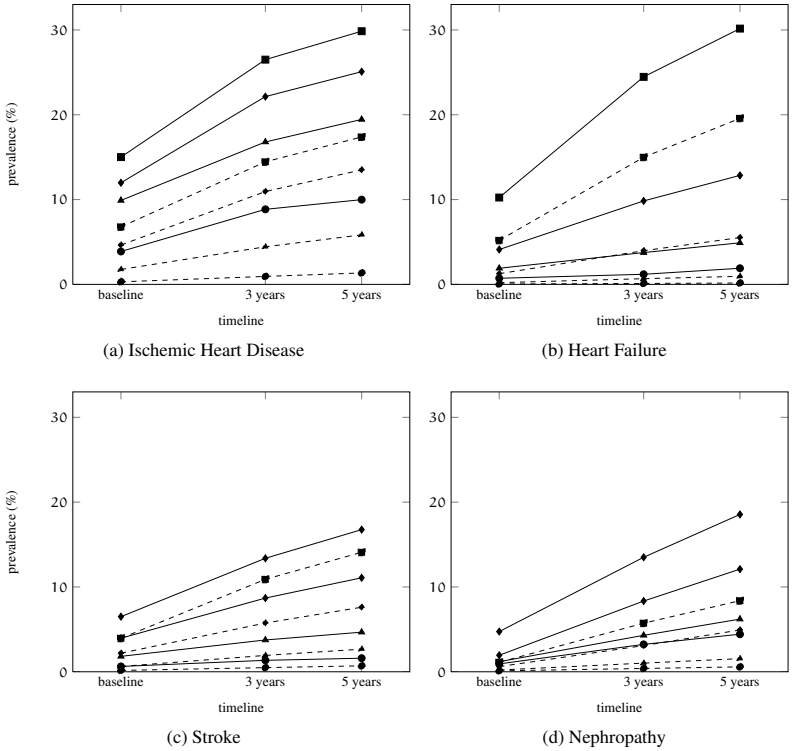
Table A.4: Base Line Prevalences and Estimated Follow-up Probabilities by Urbanity in % (with standard errors) of Health Risks and Chronic Cardiovascular Multimorbidity.

	Base Line				3 years follow-up				5 years follow-up			
Risk Factors	None	DL	HT	DL+HT	None	DL	HT	DL+HT	None	DL	HT	DL+HT
Diseases												
DM	2·3	16·7	15·8	33·1	6·0	23·9	24·0	41·5	7·5	26·2	26·4	42·8
	(0·5)	(1·2)	(1·2)	(1·5)	(0·8)	(1·3)	(1·4)	(1·6)	(0·8)	(1·4)	(1·4)	(1·6)
IHD	1·2	11·0	8·1	18·3	3·5	18·5	14·2	27·7	4·6	21·4	16·6	30·5
	(0·3)	(1·0)	(0·9)	(1·2)	(0·6)	(1·2)	(1·1)	(1·4)	(0·7)	(1·3)	(1·2)	(1·5)
HF	0·4	1·3	2·9	3·2	1·4	3·6	6·2	7·4	1·9	4·7	8·1	9·5
	(0·2)	(0·4)	(0·5)	(0·6)	(0·4)	(0·6)	(0·8)	(0·8)	(0·4)	(0·7)	(0·8)	(0·9)
ST	0·5	3·2	3·9	7·0	1·8	6·7	8·0	12·4	2·5	8·6	9·8	14·2
	(0·2)	(0·6)	(0·6)	(0·8)	(0·4)	(0·8)	(0·9)	(1·0)	(0·5)	(0·9)	(0·9)	(1·1)
NP	0·2	0·4	1·8	1·8	1·1	2·1	5·5	6·0	1·7	3·6	8·4	9·3
	(0·1)	(0·2)	(0·4)	(0·4)	(0·3)	(0·5)	(0·7)	(0·8)	(0·4)	(0·6)	(0·9)	(0·9)
RP	$<$·1	0·1	0·2	0·4	0·1	0·5	0·6	0·9	0·3	0·9	1·1	1·6
	(0·1)	(0·1)	(0·1)	(0·2)	(0·1)	(0·2)	(0·2)	(0·3)	(0·2)	(0·3)	(0·3)	(0·4)
$\geqslant 1$	4·1	28·3	27·8	50·0	11·6	41·3	43·3	65·1	14·9	45·8	49·2	69·3
	(0·6)	(1·4)	(1·4)	(1·6)	(1·0)	(1·6)	(1·6)	(1·5)	(1·1)	(1·6)	(1·6)	(1·5)

Table A.5: Probability of having a diagnosis of a chronic cardiovascular disease at baseline, and at three and five years follow-up, under condition of the presence or absence of health risks. Abbreviations: HT=hypertension; DL=dyslipidemia; DM=diabetes mellitus; IHD=ischemic heart disease; HF=heart failure; ST=stroke; NP=nephropathy; RP=retinopathy. Results are shown in percentages and standard error (SE).

	Base Line				3 years follow-up				5 years follow-up			
Risk factors	None	DL	HT	DL+HT	None	DL	HT	DL+HT	None	DL	HT	DL+HT
DM+IHD	0.20	2.62*	1.76	6.16	0.71	5.76*	4.44*	11.19	0.95	6.77*	5.38*	13.98*
	(0.14)	(0.50)	(0.41)	(0.76)	(0.27)	(0.74)	(0.65)	(0.99)	(0.31)	(0.79)	(0.71)	(1.10)
IHD+HF	0.09	0.83*	0.87*	1.77*	0.36	2.12*	2.24*	3.88*	0.55	2.80*	3.16*	5.40*
	(0.09)	(0.29)	(0.29)	(0.42)	(0.19)	(0.45)	(0.47)	(0.61)	(0.23)	(0.52)	(0.55)	(0.71)
DM+HF	0.08	0.54	0.70	1.33	0.36	1.63*	2.14*	3.89*	0.53	2.48*	2.98*	4.99*
	(0.09)	(0.23)	(0.26)	(0.36)	(0.19)	(0.40)	(0.46)	(0.61)	(0.23)	(0.49)	(0.54)	(0.69)
DM+NP	0.04	0.28	0.44	0.87	0.29	1.11*	2.04*	3.77*	0.50	1.73*	3.13*	5.04*
	(0.06)	(0.17)	(0.21)	(0.29)	(0.17)	(0.33)	(0.45)	(0.60)	(0.22)	(0.41)	(0.55)	(0.69)
HF+ST	0.03	0.10	0.33	0.36	0.16	0.51	0.91	1.43*	0.25	0.86	1.41*	2.19*
	(0.06)	(0.10)	(0.18)	(0.19)	(0.13)	(0.23)	(0.30)	(0.38)	(0.16)	(0.29)	(0.37)	(0.46)
DM+ST	0.02	0.65	0.65	2.63	0.22	1.88	2.14	5.25	0.36	2.40	2.91	6.38
	(0.05)	(0.25)	(0.25)	(0.51)	(0.15)	(0.43)	(0.46)	(0.71)	(0.19)	(0.48)	(0.53)	(0.77)
IHD+ST	0.02	0.55	0.44	1.74	0.15	1.61	1.33	3.84	0.26	2.30	1.96	4.94*
	(0.04)	(0.23)	(0.21)	(0.41)	(0.12)	(0.40)	(0.36)	(0.61)	(0.16)	(0.47)	(0.44)	(0.68)
HF+NP	0.01	0.02	0.10	0.19	0.14	0.37	0.88*	1.11*	0.27	0.59	1.53*	1.97*
	(0.04)	(0.05)	(0.10)	(0.14)	(0.12)	(0.19)	(0.30)	(0.33)	(0.16)	(0.24)	(0.39)	(0.44)
IHD+NP	0.01	0.08	0.16	0.52	0.11	0.63	1.04	2.02	0.20	1.05	1.73	3.35*
	(0.03)	(0.09)	(0.13)	(0.23)	(0.10)	(0.25)	(0.32)	(0.44)	(0.14)	(0.32)	(0.41)	(0.57)
DM+RP	0.01	0.11	0.13	0.27	0.09	0.43	0.43	0.71	0.20	0.65	0.89*	1.32*
	(0.04)	(0.10)	(0.11)	(0.16)	(0.09)	(0.20)	(0.21)	(0.26)	(0.14)	(0.25)	(0.30)	(0.36)

Table A.6: Probability of having comorbid combinations of chronic cardiovascular diseases at baseline, and 3 and 5 years follow-up, under condition of the presence or absence of health risks. Abbreviations: HT=hypertension; DL=dyslipidemia; DM=diabetes mellitus; IHD=ischemic heart disease; HF=heart failure; ST=stroke; NP=nephropathy; RP=retinopathy. Results are shown in percentages and standard error (SE). Incidence rates that deviate significantly ($p < 0.001$) from expected values based on the individual rates, and have a clinical importance as well (absolute increase $> 0.5\%$), are denoted with a *.



(a) Ischemic Heart Disease

(b) Heart Failure

(c) Stroke

(d) Nephropathy

Figure A.3: Prevalences of ischemic heart disease, heart failure, stroke, and nephropathy for diabetics and non-diabetics, at baseline, and 3 and 5 years follow-up. There are four age-groups, 35-50 years, 50-65 years, 65-80 years, and above 80 years, which are respectively represented by circles, triangles, diamonds, and squares. Straight lines represent prevalences for diabetics and the dotted lines represent prevalences for non-diabetics.

BIBLIOGRAPHY

[1] S. Acid, L. de Campos, J. Fernández-Luna, S. Rodríguez, J. Rodríguez, and J. Salcedo. A comparison of learning algorithms for bayesian networks: a case study based on data from an emergency medical service. *Artificial Intelligence in Medicine*, 30:215–232, 2004.

[2] H. Akaike. A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, 19(6):716–723, 1974.

[3] A. Alekseyenko, N. I. Lytkin, J. Ai, B. Ding, L. Padyukov, C. F. Aliferis, and A. Statnikov. Causal graph-based analysis of genome-wide association data in rheumatoid arthritis. *Biology Direct*, 6:25–37, 2011.

[4] S. Amstrup, E. DeWeaver, D. Douglas, B. Marcot, G. Durner, C. Bitz, and D. Bailey. Greenhouse gas mitigation can reduce sea-ice loss and increase polar bear persistence. *Nature*, 468:955–958, 2010.

[5] S. Andersson, D. Madigan, and M. Perlman. An alternative markov property for chain graphs. *Scand. J. Statist*, 28:33–85, 1996.

[6] L. Andrade, I. B. nor, M. Viana, S. Andreoni, and Y. Wang. Clustering of psychiatric and somatic illnesses in the general population: multimorbidity and socioeconomic correlates. *Brazilian Journal of Medical and Biological Research*, 43:483–491, 2010.

[7] A. Angold, E. Costello, and A. Erkanli. Comorbidity. *J Child Psychol Psychiatry*, 40(1):57–87, 1999.

[8] A. Aussem, S. de Morias, and M. Corbex. Analysis of nasopharyngeal carcinoma risk factors with Bayesian networks. *Artificial intelligence in Medicine*, 54(1):53–62, 2012.

[9] P. Austin, V. Goel, and C. Walraven. An introduction to multilevel regression models. *Can J Public Health*, 92(2):150–4, 2001.

[10] A. L. Barabási, N. Gulbahce, and J. Loscalzo. Network medicine: a network-based approach to human disease. *Nat Rev Genet*, 12(1):56–68, 2011.

[11] K. Barnett, S. Mercer, M. Norbury, G. Watt, S. Wyke, and B. Guthrie. Epidemiology of multimorbidity and implications for health care, research, and medical education: a cross-sectional study. *Lancet*, 380:37–43, 2012.

[12] D. Bartholomew. *Latent variable models and factor analysis*. Oxford University Press, New York, 1987.

[13] L. Batstra, E. Bos, and J. Neeleman. Quantifying psychiatric comorbidity, lessons from chronic disease epidemiology. *Soc Psychiatry Psychiatr Epidemiol*, 37:105–111, 2002.

[14] K. Beuker, S. Schjolberg, K. Lie, R. Donders, M. Lappenschaar, S. Swinkels, and J. Buitelaar. The structure of autism spectrum disorder symptoms in the general population at 18 months. *Journal of Autism and Developmental Disorders*, 43(1):45–56, 2013.

[15] A. Bocquier, S. Cortaredona, S. Nauleau, M. Jardin, and P. Verger. Prevalence of treated diabetes: Geographical variations at the small-area level and their association with area-level characteristics. a multilevel analysis in southeastern france. *Diabetes and Metabolism*, 37(1):39–46, 2011.

[16] C. Boyd and M. Fortin. Future of multimorbidity research: How should understanding of multimorbidity inform health system design? *Public Health Reviews*, 32(2):451–474, 2005.

[17] C. Boyd, J. Darer, C. Boult, L. Fried, L. Boult, and A. Wu. Clinical practice guidelines and quality of care for older patients, with multiple comorbid diseases: Implications for pay for performance. *JAMA*, 294(6):716–724, 2005.

[18] N. Breslow and D. Clayton. Approximate inference in generalized linear mixed models. *Journal of Statistical Computation and Simulation*, 88:9–25, 1993.

[19] H. Britt, C. Harrison, G. Miller, and S. Knox. Prevalence and patterns of multimorbidity in australia. *Med J Aust*, 189(2):72–77, 2008.

[20] W. Brown, D. Draper, H. Goldstein, and J. Rasbash. Bayesian and likelihood methods for fitting multilevel models with complex level-1 variation. *Computational Statistics and Data Analysis*, 39:203–225, 2002.

[21] W. Browne and D. Draper. A comparison of Bayesian and likelihood-based methods for fitting multilevel models. *Bayesian Analysis*, 1(3):473–514, 2006.

[22] K. Burnham and D. Anderson. Multimodel inference – understanding AIC and BIC in model selection. *Sociological Methods and Research*, 33(2):261–304, 2004.

[23] D. Byar. Why data bases should not replace randomized clinical trials. *Biometrics*, 36:337–342, 1980.

[24] C. P. D. Campos and F. G. Cozman. Belief updating and learning in semi-qualitative probabilistic networks. In *UAI*, pages 153–160, 2005.

[25] O. Carretero and S. Oparil. Essential hypertension – part I: Definition and etiology. *Circulation*, 101:329–335, 2000.

[26] C. Caughey, A. Vitry, A. Gilbert, and E. Roughead. Prevalence of comorbidity of chronic diseases in australia. *BMC Public Health*, 8:221, 2008.

[27] M. Charlson, K. Ales, P. Pompei, and K. MacKenzie. A new method of classifying prognostic comorbidity in longitudinal studies. *J Chronic Dis*, 40, 1987.

[28] L. Chen, N. Blumm, N. Christakis, A. Barabasi, and T. Deisboeck. Cancer metastasis networks and the prediction of progression patterns. *British journal of Cancer*, 101:749–758, 2009.

[29] T. Claassen. *Causal Discovery and Logic*. PhD thesis, Radboud University Nijmegen, Nijmegen, The Netherlands, 2013.

[30] P. Cleary, S. Greenfield, H. Mulley, S. Pauker, S. Schroeder, L. Wexler, and B. McNeil. Cancer metastasis networks and the prediction of progression patterns. *JAMA*, 266:73–79, 1991.

[31] G. F. Cooper, C. F. Aliferis, R. Ambrosino, J. Aronis, B. G. Buchanan, R. Caruana, M. J. Fine, C. Glymour, G. Gordon, B. H. Hanusa, J. E. Janosky, C. Meek, T. Mitchell, T. Richardson, and P. Spirtes. An evaluation of machine-learning methods for predicting pneumonia mortality. *Artificial Intelligence in Medicine*, 9(2):107–138, 1997.

[32] R. Cowell, A. Dawid, S. Lauritzen, and D. Spiegelhalter. *Probabilistic Networks and Expert Systems*. Springer, New York, USA, 1999.

[33] A. Cramer, L. Waldorp, H. van der Maas, and D. Borsboom. Comorbidity: a network perspective. *Behavioral and brain sciences*, 33:137–193, 2010.

[34] A. Dawid. Beware of the dag! In *JMLR: Workshop and Conference Proceedings*, volume 6, pages 69–86, 2009.

[35] C. P. de Campos, L. Zhang, Y. Tong, and Q. Ji. Semi-qualitative probabilistic networks in computer vision problems. *Journal of Statistical Theory and Practice*, 3(1):197–210, 2009.

[36] M. de Jongh and M. Druzdzel. A comparison of structural distance measures for causal Bayesian network models. In *Recent Advances in Intelligent Information Systems*, pages 443–456, Warsaw, Poland, 2009. Academic Publishing House EXIT.

[37] S. de Lusignan and C. van Weel. The use of routinely collected computer data for research in primary care: opportunities and challenges. *Family Practice*, 23: 253–263, 2006.

[38] L. Deckx, M. van den Akker, J. Metsemakers, A. Knottnerus, F. Schellevis, and F. Buntinx. Chronic diseases among older cancer survivors. *Journal of Cancer Epidemiology*, 2012:Article ID 206414, 7 pages, 2012.

[39] C. Diederichs, K. Berger, and D. Bartels. The measurement of multiple chronic diseases - a systematic review on existing multimorbidity indices. *J Gerontol A Biol Sci Med Sci*, 66(3):301–311, 2011.

[40] K. Diehl and S. Schneider. How relevant are district characteristics in explaining subjective health in Germany? a multilevel analysis. *Social Science and Medicine*, 72(7):1205–1210, 2011.

[41] H. Drooglever-Fortuyn, M. Lappenschaar, F. Nienhuis, J. Furer, P. Hodiamont, C. Rijnders, G. Lammers, W. Renier, J. Buitelaar, and S. Overeem. Psychotic symptoms in narcolepsy: phenomenology and a comparison with schizophrenia. *General Hospital Psychiatry*, 31(2):146–154, 2009.

[42] H. Drooglever-Fortuyn, M. Lappenschaar, J. Furer, P. Hodiamont, C. Rijnders, W. Renier, J. Buitelaar, and S. Overeem. Anxiety and mood disorders in narcolepsy: a case–control study. *General Hospital Psychiatry*, 32(1):49–56, 2010.

[43] H. Drooglever-Fortuyn, R. Fonczek, M. Smitshoek, S. Overeem, M. Lappenschaar, J. Kalkman, W. Renier, J. Buitelaar, G. Lammers, and G. Blijenberg. Severe fatigue in narcolepsy with cataplexy. *Journal of Sleep Research*, 21(2): 163–169, 2012.

[44] M. Druzdzel. *Probabilistic Reasoning in Decision Support Systems: From Computation to Common Sense*. PhD thesis, Carnegie-Mellon University, Pittsburgh, Pennsylvania, 1993.

[45] M. Druzdzel and M. Henrion. Efficient reasoning in qualitative probabilistic networks. In *Proceedings of the Eleventh National Conference on Artificial Intelligence*, pages 548–553, California, CA, USA, 1993. AAAI Press.

[46] M. Druzdzel and L. van der Gaag. Elicitation of probabilities for belief networks: Combining qualitative and quantitative information. In P. Besnard and S. Hanks, editors, *Proceedings of the Eleventh conference on Uncertainty in artificial intelligence*, pages 141–148, Montreal, Quebec, Canada, 1995. Morgan Kaufmann.

[47] M. J. Druzdzel and M. Henrion. Intercausal reasoning with uninstantiated ancestor nodes. In *UAI*, pages 317–325, 1993.

[48] J. Elhai, P. Calhoun, and J. Ford. Statistical procedures for analyzing mental health services data. *Psychiatry Research*, 160:129–136, 2008.

[49] E. Elm, D. Altman, M. Egger, S. Pocock, P. Gotzsche, and J. Vandenbroucke. The strengthening the reporting of observational studies in epidemiology (strobe) statement: Guidelines for reporting observational studies. *PLoS Medicine*, 4:10–15, 2008.

[50] A. Feinstein. The pretherapeutic classification of comorbidity in chronic disease. *J Chronic Dis*, 23:455–468, 1970.

[51] M. Flores, A. Nicholson, A. Burnskill, K. Korb, and S. Mascaro. Incorporating expert knowledge when learning Bayesian network structure: A medical case study. *Artificial Intelligence in Medicine*, 53(3):181–204, 2011.

[52] M. Fortin, C. Hudon, L. Lapointe, and M. van A Vanassel. Multimorbidity is common to family practice: Is it commonly researched? *Canadian Family Physician*, 51:245, 2005.

[53] M. Fortin, C. Hudon, J. Haggerty, M. van den Akker, and J. Almirall. Prevalence estimates of multimorbidity: a comparative study of two sources. *BMC Health Services Research*, 10:111, 2010.

[54] M. Fortin, M. S. adn ME Poitras, J. Almirall, and H. Maddocks. A systematic review of prevalence studies on multimorbidity: toward a more uniform methodology. *Ann Fam Med*, 10(2):142–151, 2012.

[55] E. France, S. Wyke, J. Gunn, F. Mair, G. McLean, and S. Mercer. Multimorbidity in primary care: a systematic review of prospective cohort studies. *Br J Gen Pract*, 62(597):e297–307, 2012.

[56] N. Friedman, M. Goldszmidt, and A. Wyner. Data analysis with Bayesian networks: A bootstrap approach. In K. Laskey and H. Prade, editors, *Proceedings of the 15th Annual Conference on Uncertainty in Artificial Intelligence (UAI-99)*, pages 206–215, Stockholm, Sweden, 1999. Morgan Kaufmann.

[57] M. Frydenberg. The chain graph Markov property. *Scand J Statist*, 17:333–353, 1990.

[58] M. Fulton and E. Allen. Polypharmacy in the elderly: A literature review. *Journal of the American Acadamy of Nurse Practitioners*, 17(4):123–132, 2005.

[59] G. Galilei. *Discorsi e dimostrazioni matematiche, intorno Ã due nuove scienze*. Elzevier, Leiden, 1638.

[60] D. Geiger, T. Verma, and J. Pearl. d-separation: From theorems to algorithms. In *Proceedings of the Fifth Conference Annual Conference on Uncertainty in Artificial Intelligence (UAI-89)*, pages 118–125, Corvallis, Oregon, 1989. AUAI Press.

[61] L. Glynn, B. Buckley, D. Reddan, J. Newell, J. Hinde, S. Dinneen, and A. Murphy. Multimorbidity and risk among patients with established cardiovascular disease. *British Journal of General Practice*, 58:488–494, 2008.

[62] H. Goldstein. Multilevel mixed linear model analysis using iterative generalized least squares. *Biometrika*, 73(1):43–56, 1986.

[63] H. Goldstein. Restricted unbiased iterative generalised least squares estimation. *Biometrika*, 76:622–623, 1989.

[64] H. Goldstein. Heteroscedasticity and complex variation. *Encyclopedia of Statistics in Behavioral Science*, 2:790–795, 2005.

[65] H. Goldstein and J. Rabash. Improved approximations for multilevel models with binary responses. *Journal of the Royal Statistical Society (Series A)*, 159: 505–512, 1996.

[66] H. Goldstein, W. Browne, and J. Rasbash. Multilevel modelling of medical data. *Statistics in Medicine*, 21(21):3291–3315, 2002.

[67] T. Gress, F. Nieto, E. Shahar, M. Wofford, and F. Brancati. Hypertension and antihypertensive therapy as risk factors for type 2 diabetes mellitus. *New England Journal of Medicine*, 342(13):905–912, 2000.

[68] J. Gunn, D. Ayton, K. Densley, J. Pallant, P. Chondros, H. Herrman, and C. Dowrick. The association between chronic illness, multimorbidity and depressive symptoms in an australian primary care cohort. *Soc Psychiatry Psychiatr Epidemiol*, 47:175–184, 2012.

[69] B. Guthrie, K. Payne, P. Alderson, M. McMurdo, and S. Mercer. Adapting clinical guidelines to take account of multimorbidity. *BMJ*, 345:e6341, 2012.

[70] E. Gyftodimos and P. Flach. Hierarchical Bayesian networks: An approach to classification and learning for structured data. In G. Vouros and T. Panayiotopoulos, editors, *Methods and Applications of Artificial Intelligence*, volume 3025 of *Lecture Notes in Computer Science*, pages 291–300, Samos, Greece, 2004. Springer.

[71] S. Hall. A user's guide to selecting a comorbidity index for clinical research. *Journal of Clinical Epidemiology*, 59:849–855, 2006.

[72] D. Heckerman and B. Nathwani. Toward normative expert systems: part I – the pathfinder project. *Methods of Information in Medicine*, 31:90–105, 1992.

[73] D. Heckerman, E. Horvitz, and B. Nathwani. Toward normative expert systems: part II – probability-based representations for efficient knowledge acquisition and inference. *Methods of Information in Medicine,*, 31:106–116, 1992.

[74] D. Heckerman, D. Geiger, and D. Chickering. Learning Bayesian networks: The combination of knowledge and statistical data. *Machine Learning*, 20(3): 197–243, 1995.

[75] G. Henriksson, G. Weitoft, and P. Allebeck. Associations between income inequality at municipality level and health depend on context - a multilevel analysis on myocardial infarction in Sweden. *Social Sciences and Medicine*, 71(6): 1141–1149, 2010.

[76] M. Henrion and M. Druzdzel. Qualitative propagation and scenario-based approaches to explanation in probabilistic reasoning. *Uncertainty in Artificial Intelligence*, 6:17–32, 1991.

[77] J. Hippisley-Cox, V. Hammersley, M. Pringle, C. Coupland, N. Crown, and L. Wright. Methodology for assessing the usefullness of general practice data for research in one research network. *Health Informatics Journal*, 10(2):91–109, 2004.

[78] K. Ho, J. Pinsky, W. Kannel, and D. Levy. The epidemiology of heartfailure: The framingham study. *Journal of the American College of Cardiology*, 22(4): 6–13, 1993.

[79] N. Hoeymans and F. Schellevis. Selection of chronic diseases. Technical report, National Kompass Public Health - National Institute for Public Health and the Enviroment, the Netherlands, 2012.

[80] A. Hofman, C. van Duijn, O. Franco, M. Ikram, H. Janssen, C. Klaver, E. Kuipers, T. Nijsten, B. Stricker, H. Tiemeier, A. Uitterlinden, M. Vernooij, and J. Witteman. The rotterdam study: 2012 objectives and design update. *Eur J Epidemiol*, 26:657–686, 2012.

[81] L. Holden, P. Scuffham, M. Hilton, A. Muspratt, S. Ng, and H. Whiteford. Patterns of multimorbidity in working australians. *Population Health Metrics*, 9(1): 115, 2011.

[82] A. Hommersom. *On the Application of Formal Methods to Clinical Guidelines*. PhD thesis, Radboud University Nijmegen, Nijmegen, The Netherlands, 2008.

[83] J. Hox. *Multilevel Analysis: techniques and applications*. Routledge, New York, USA, 2010.

[84] L. Huges, M. McMurdo, and B. Guthrie. Guidelines for people not for diseases: the challenges of applying uk clinical guidelines to people with multimorbidity. *Age and Ageing*, 42(1):62–69, 2013.

[85] A. Huntley, R. Johnson, S. Purdy, J. Valderas, and C. Salisbury. Measures of multimorbidity and morbidity burden for use in primary care and community settings: a systematic review and guide. *Ann Fam Med*, 10(2):134–141, 2012.

[86] J. Jaccard and R. Turrisi. Interaction effects in multiple regression. *Sage University Papers Series on Quantitative Applications in the Social Sciences*, pages 07–072, 2003. Thousand Oaks, CA: Sage.

[87] R. Jackson, C. Lawes, D. Bennet, R. Milne, and A. Rodgers. Treatment with drugs to lower bloodpressure and blood cholesterol based on an individual's absolute cardiovascular risk. *Lancet*, 365:434–441, 2005.

[88] X. Jiang, M. Barmada, G. Cooper, and M. Becich. A Bayesian method for evaluating and discovering disease loci associations. *PLoS One*, 6(8):e22075, 2011.

[89] X. Jiang, R. Neapolitan, M. Barmada, and S. Viswesaran. Learning genetic epistasis using Bayesian network scoring criteria. *BMC Bioinformatics*, 12:89, 2011.

[90] R. Johnson and F. Wolinksy. The structure of health status among older adults: disease, disability, functional limitation, and perceived health. *Journal of Health and Social Behavior*, 34:105–121, 1993.

[91] M. Kaplan and A. Feinstein. The importance of classifying initial comorbidity in evaluating the outcome of diabetes mellitus. *J Chronic Dis*, 27(7-8):387–404, 1974.

[92] I. Kirchberger, C. Meisinger, M. Heier, A. Zimmerman, B. Thorand, C. Autenrieth, A. Peters, K. Ladwig, and A. Döring. Patterns of multimorbidity in the aged population - results from the kora-age study. *PLoS One*, 7(1):e30556, 2012.

[93] D. Koller and N. Friedman. *Probabilistic Graphical Models*. The MIT Press, Cambridge, Massachusetts, London, 2009.

[94] D. Koller and A. Pfeffer. Object-oriented Bayesian networks. In D. Geiger, P. Prakash, and P. Shenoy, editors, *Proceedings of the Thirteenth Conference on Uncertainty in Artificial Intelligence*, pages 302–313, Providence, Rhode Island, USA, 1997. Morgan Kaufmann.

[95] A. Kolmogorov and S. Fomin. *Elements of the Theory of Functions and Functional Analysis*. Graylock Press, New York, USA, 1957.

[96] E. Korendijk, C. Maas, M. Moerbeek, and P. van der Heijden. The influence of misspecification of the heteroscedasticity on multilevel regression parameter and standard error estimates. *Methodology*, 2(4):67–72, 2008.

[97] M. Korver and P. Lucas. Converting a rule-based expert system into a belief network. *Medical Informatics*, 18(3):219–241, 1993.

[98] H. Kraemer. Statistical issues in assessing comorbidity. *Statistics in Medicine*, 14:721–733, 1995.

[99] S. Kullback and R. Leibler. On information and sufficiency. *Annals of Mathematical Statistics*, 22:79–86, 1951.

[100] U. Kumari and K. Heese. Cardiovascular dementia - a different perspective. *The Open Biochemistry Journal*, 4:29–52, 2010.

[101] J. Lagro. *Cardiovascular and cerebrovascular physiological measurements in clinical practice and prognostics in geriatric patients*. PhD thesis, Radboud University Nijmegen, Nijmegen, The Netherlands, 2013.

[102] M. Lappenschaar, A. Hommersom, and P. Lucas. Probabilistic causal models of multimorbidity concepts. In *AMIA Proceedings of the 2012 Annual Symposium*, pages 475–484, Chicago, United States, November 2012.

[103] M. Lappenschaar, A. Hommersom, J. lagro, and P. Lucas. Understanding the co-occurrence of diseases using structure learning. In *Artificial Intelligence in Medicine, Lecture Notes in Computer Science*, volume 7885, pages 135–144, 2013.

[104] M. Lappenschaar, A. Hommersom, and P. Lucas. Qualitative chain graphs and their application. *Journal of Approximate Reasoning*, 2013. under revision.

[105] M. Lappenschaar, A. Hommersom, P. Lucas, J. Lagro, and S. Visscher. Multilevel Bayesian networks for the analysis of hierarchical health data. *Artificial Intelligence in Medicine*, 57:171–183, 2013.

[106] M. Lappenschaar, A. Hommersom, P. Lucas, J. Lagro, S. Visscher, J. Korevaar, and F. Schellevis. Multilevel temporal bayesian networks can model longitudinal change in multimorbidity. *Journal of Clinical Epidemiology*, 66:1405–1416, 2013.

[107] S. Lauritzen and T. Richardson. Chain graph models and their causal interpretations. *J.R. Statist. Soc. B*, 64(3):321–361, 2002.

[108] S. Lauritzen and N. Wermuth. Graphical models for associations between variables, some of which are qualitative and some quantitative. *The Annals of Statistics*, 17(1):31–57, 1989.

[109] U. Lerner, E. Segal, and D. Koller. Exact inference in networks with discrete children of continuous parents. In J. Breese and D. Koller, editors, *Proceedings of the 17th Conference in Uncertainty in Artificial Intelligence*, pages 319–328, San Francisco, CA, USA, 2001. Morgan Kaufmann.

[110] X.-K. Li and S.-Z. Liao. Inference in qpns with multiple observations. In *Machine Learning and Cybernetics, 2009 International Conference on*, volume 1, pages 221 –226, july 2009.

[111] L. Lim, E. Lamoureux, S. Saw, W. Tay, P. Mitchell, and T. Wong. Are myopic eyes less likely to have diabetic retinopathy? *Ophthalmology*, 117(3):524–530, 2010.

[112] E. Lin, S. Heckbert, C. Rutter, P. Ciechanowski, E. Ludman, M. Oliver, B. Young, D. McCulloch, and M. von Korff. Depression and increased mortality in diabetes: unexpected causes of death. *Ann Fam Med*, 7:414–421, 2009.

[113] B. Linn, M. Linn, and G. Lee. Cumulative illness rating scale. *J Am Geriatr Soc*, 5:622–626, 1968.

[114] C.-L. Liu and M. P. Wellman. Bounding probabilistic relationships in bayesian networks using qualitative influences: methods and applications. *Int. J. Approx. Reasoning*, 36(1):31–73, 2004.

[115] Z. Liu, B. Malone, and C. Yuan. Empirical evaluation of scoring functions for Bayesian network model selection. *BMC Bioinformatics*, 13(Suppl 15):S16, 2012.

[116] P. Lucas. Bayesian analysis, pattern analysis, and data mining in health care. *Current Opinion in Critical Care*, 10:399–403, 2004.

[117] P. Lucas. Bayesian network modelling through qualitative patterns. *Artificial Intelligence*, 163(2):233–263, 2005.

[118] P. Lucas, H. Boot, and B. Taal. Computer-based decision-support in the management of primary gastric non-hodgkin lymphoma. *Methods of Information in Medicine*, 37:206–219, 1998.

[119] P. Lucas, N. de Bruijn, K. Schurink, and I. Hoepelman. A probabilistic and decision-theoretic approach to the management of infectious disease at the icu. *Artificial Intelligence in Medicine*, 19(3):251–279, 2000.

[120] P. Lucas, L. van der Gaag, and A. Abu-Hanna. Bayesian networks in biomedicine and health-care. *Artificial Intelligence in Medicine*, 30(3):201–214, 2004.

[121] M. Madden. On the classification performance of TAN and general Bayesian networks. *Knowledge Based Systems*, 22(7):489–495, 2009.

[122] C. Magri, N. Calleja, C. Buhagiar, S. Fava, and J. Vassallo. Factors associated with diabetic nephropathy in subjects with proliferative retinopathy. *Int Urul Nephrol*, 44:197–206, 2012.

[123] K. Manton and M. Woodbury. A new procedure for analysis of medical classification. *Methods of Information in Medicine*, 21, 1982.

[124] A. Marengoni, D. Rizzuto, H. Wang, B. Winblad, and L. Fratiglioni. Patterns of chronic multimorbidity in the elderly population. *J Am Geriatr Soc*, 57:225–230, 2009.

[125] A. Marengoni, S. Angleman, R. Melis, F. Mangialasche, A. Karp, A. Garmen, B. Meinow, and L. Fratiglioni. Aging with multimorbidity: A systematic review of the literature. *Ageing Research Reviews*, 10:430–439, 2011.

[126] D. Margaritis. *Learning Bayesian Network Model Structure from Data*. PhD thesis, School of Computer Science, Carnegie-Mellon University, Pittsburgh, PA, 2003. Available as Technical Report CMU-CS-03-153.

[127] A. B. Mariotto, J. H. Rowland, L. A. Ries, S. Scoppa, and E. J. Feuer. Multiple cancer prevalence: A growing challenge in long-term survivorship. *Cancer Epidemiology, Biomarkers and Prevention*, 16:566–571, 2007.

[128] E. Marshall and D. Spiegelhalter. Approximate cross-validatory predictive checks in disease mapping models. *Statistics in Medicine*, 22:1649–1660, 2003.

[129] J. McMurray, S. Adamapoulos, S. Anker, A. Auricchio, M. Böhm, K. Dickstein, V. Falk, G. Filippatos, C. F. M. Gomez-Sanchez, T. Jaarsma, L. Køber, G. Lip, A. Maggioni, A. Parkhomenko, B. Pieske, B. Popescu, P. Rønnevik, F. Rutten, J. Schwitter, P. Seferovic, J. Stepinska, P. Trindade, A. Voors, F. Zannad, and A. Zeiher. European society of cardiology guidelines for the diagnosis and treatment of acute and chronic heart failure. *European Heart Journal*, 33:1787–1847, 2012.

[130] Medical Research Council. Streptomycin treatment of pulmonary tuberculosis. *BMJ*, 2((4582):769–82, 1948.

[131] M. Meldrum. A brief history of the randomize controlled trial; from oranges and lemons to the gold standard. *Hematology/Oncology Clinics of North America*, 14(4):745–760, 2000.

[132] B. Middleton, M. Shwe, D. Heckerman, M. Henrion, E. Horvitz, H. Lehmann, and G. Cooper. Probabilistic diagnosis using a reformulation of the INTERNIST-1/QMR knowledge base, II – evaluation of diagnostic performance. *Methods of Information in Medicine*, 30:256–267, 1991.

[133] A. Mitnitski, J. Graham, A. Mogilner, and K. Rockwood. Frailty, fitness and late-life mortality in relation to chronological and biological age. *BMC Geriatrics*, 2:1, 2002.

[134] R. Moineddin, F. Matheson, and R. Glazier. A simulation study of sample size for multilevel logistic regression models. *BMC Medical Research Methodology*, 7(34), 2007.

[135] R. Moretti, P. Torre, R. Antonello, D. Manganaro, C. Vilotti, and G. Pizzolato. Risk factors for vascular dementia: Hypotension as a key point. *Vascular Health Risk Management*, 4(2):395–402, 2008.

[136] E. Moricke, M. Lappenschaar, S. Swinkels, N. Rommelse, and J. Buitelaar. Latent class analysis reveals five homogeneous behavioural and developmental profiles in a large dutch population sample of infants aged 14–15 months. *European Child and Adolescent Psychiatry*, 22(2):103–115, 2013.

[137] A. Ng and M. Jordan. On discriminative vs. generative classifiers: A comparison of logistic regression and naive bayes, 2002.

[138] G. Ng and K. Ong. Using a qualitative probabilistic network to explain diagnostic reasoning in an expert system for chest pain diagnosis. In *Computers in Cardiology 2000*, pages 569–572. IEEE, 2000.

[139] S. ng, L. Holden, and J. Sun. Indentifying comorbidity patterns of health conditions via cluster analysis of pairwise concordance statistics. *Statistics in Medicine*, 31:3393–3405, 2012.

[140] M. Nielen, F. Schellevis, and R. Verheij. Inter-practice variation in diagnosing hypertension and diabetes mellitus: a cross-sectional study in general practice. *BMC Family Practice*, 10:1–6, 2009.

[141] M. Nielen, A. van Sijl, M. Peters, R. Verheij, F. Schellevis, and M. Nurmohamed. Cardiovascular disease prevalence in patients with inflammatory arthritis, diabetes mellitus and osteoarthritis: a cross-sectional study in primary care. *BMC Musculoskeletal Disorders*, 13:150–155, 2012.

[142] J. Nuyen, A. Volkers, P. Verhaak, F. Schellevis, P. Groenewegen, and G. van den Bos. Accuracy of diagnosing depression in primary care: the impact of chronic somatic and psychiatric co-morbidity. *Psychological Medicine*, 35:1185–1195, 2005.

[143] K. Nylund, T. Asparouhov, and B. Muthen. Deciding on the number of classes in latent class analysis and growth mixture modeling. a monte carlo simulation study. *Structural Equation Modeling*, 14:535–569, 2007.

[144] J. O'Halloran, G. Miller, and H. Britt. Defining chronic conditions for primary care with icpc-2. *Familiy Practice*, 21:381–386, 2004.

[145] K. Olesen and S. Andreassen. Specification of models in large expert systems based on causal probabilistic networks. *Artificial Intelligence in Medicine*, 5(3): 269–281, 1993.

[146] A. Oniśko, M. Druzdzel, and H. Wasyluk. Extension of the hepar II model to multiple-order diagnosis. In *Intelligent Information Systems - Advances in Soft Computing Series*, pages 303–313. Springer-Verlag, Heidelberg, Germany, 2000.

[147] A. Padros-Torres, B. Poblador-Plou, A. Calderon-Larranaga, L. Gimeno-Feliu, F. Gonzalez-Rubio, A. Poncel-Falco, A. Sicras-Mainar, and J. Alcala-Nalvaiz. Multimorbidity patterns in primary care: Interactions among chronic diseases using factor analysis. *PLoS One*, 7:e32190–0, 2012.

[148] M. Paul, S. Andreassen, A. Nielsen, E. Tacconelli, N. Almanasreh, A. Fraser, D. Yahav, R. Ram, and L. Leibovici. Prediction of bacteremia using treat, a computerized decision-support system. *Clinical Infectious Diseases*, 42:1274–1282, 2006.

[149] J. Pearl. *Probabilistic Reasoning in Intelligent Systems*. Morgan Kaufmann, San Francisco, CA, USA, 1988.

[150] J. Pearl. Causal Diagrams for Empirical Research. *Biometrika*, 82(4):669–688, 1995.

[151] J. Pearl. *Causality: Models, Reasoning and Inference*. MIT press, Cambridge, Massachusetts, USA, 2000.

[152] C. Peirce and J. Jastrow. On small differences in sensation. *Memoirs of the National Academy of Sciences*, 3:73–83, 1885.

[153] J. Peña. Faithfulness in chain graphs: The discrete case. *International Journal of Approximate Reasoning*, 50(8):1306–1313, 2009.

[154] J. Peña. Faithfulness in chain graphs: The gaussian case. In *Proceedings of the 14th International Conference on Artificial Intelligence and Statistics (AISTATS 2011)*, pages 588–599, 2011.

[155] M. Pencina, R. D. Sr., R. D. Jr., and R. Vasan. Evaluating the added predictive ability of a new marker: From area under the roc curve to reclassification and beyond. *Statistics in Medicine*, 27:157–172, 2008.

[156] J. Perk, G. de Backer, H. Gohlke, I. Graham, Z. Reiner, W. M. Verschuren, C. Albus, P. Benlian, G. Boysen, R. Cifkova, C. Deaton, S. Ebrahim, M. Fisher, G. Germano, R. Hobbs, A. Hoes, S. Karadeniz, A. Mezzani, E. Prescott, L. Ryden, M. Schrerer, M. Syvänne, W. S. O. Reimer, C. Vrints, D. Wood, J. Zamorano, and F. Zannad. European society of cardiology guidelines on cardiovascular disease prevention in clinical practice. *European Heart Journal*, 33: 1635–1701, 2012.

[157] A. Perruccio, J. Power, and E. Badley. The relative impact of 13 chronic conditions across three different outcomes. *J Epidemiol Community Health*, 61: 1056–1061, 2007.

[158] R. Picard and D. Cook. Cross-validation of regression models. *Journal of the American Statistical Association*, 79(387):575–583, 1984.

[159] J. Piette and E. Kerr. The impact of comorbid chronic conditions on diabetes care. *Diabetes Care*, 29(3):725–731, 2006.

[160] M. Plummer. Penalized loss functions for Bayesian model comparison. *Biostatistics*, pages 1–17, 2008.

[161] F. Portrait, M. Lindeboom, and D. Deeg. Health and mortality of the elderly: the grade of membership method, classification and determination. *Health Econ*, 8: 441–457, 1999.

[162] M. Price, N. Welton, and A. Ades. Parameterization of treatment effects for meta-analysis in multi-state Markov models. *Statistics in Medicine*, 30:140–151, 2011.

[163] H. Putter, M. Fiocco, and R. Geskus. Parameterization of treatment effects for meta-analysis in multi-state Markov models. *Statistics in Medicine*, 26:2389–2430, 2007.

[164] S. Renooij. *Qualitative Approaches to Quantifying Probabilistic Network*. PhD thesis, Utrecht University, 2001.

[165] S. Renooij, L. C. van der Gaag, and S. Parsons. Propagation of multiple observations in qpns revisited. In F. van Harmelen, editor, *ECAI*, pages 665–669. IOS Press, 2002.

[166] J. L. Reste, P. Nabbe, B. Manceau, C. Lygidakis, C. Doerr, H. Lingner, S. Czachowski, M. Munoz, S. Argyriadou, A. Claveria, B. L. Floch, M. Barais, P. Bower, H. V. Marwijk, P. V. Ryouen, and C. Lietard. The European General Practice Research Network present a comprehensive definition of multimorbidity in family medicine and long term care, following a systematic review of relevant literature. *JAMDA*, in press, 2013.

[167] N. Rice and A. Leyland. Multilevel models: applications to health data. *J Health Services Res Pol*, 1(3):154–164, 1996.

[168] F. Richards. A flexible growth function for empirical use. . *J. Exp. Bot.*, 10: 290–300, 1959.

[169] T. Richardson. Markov properties for acyclic directed mixed graphs. *Scandinavian Journal of Statistics*, 30(1):145–157, 2003.

[170] C. S. Ritchie, E. Kvale, and M. J. Fisch. Multimorbidity: An issue of growing importance for oncologists. *Journal of Oncology Practice*, 7(6):371–374, 2011.

[171] R. Robinson. Counting unlabeled acyclic graphs. In *LNM 622*, pages 220–227. Springer, NY, 1977.

[172] K. Rockwood and A. Mitnitski. Frailty in relation to the accumulation of deficits. *Journal of Gerontology: Medical Sciences*, 62A(7):722–727, 2007.

[173] J. Rong, C. Yu, P. Yang, and J. Chen. Diabetic retinopathy: A predictor of coronary artery disease. *Diabetes and Vascular Disease Research*, online first: 1–8, 2012.

[174] F. Roque, P. Jensen, H. Schmock, M. Dalgaard, M. Andreatta, T. Hansen, K. Soeby, S. Bredkjaer, A. Juul, T. Werge, L. Jensen, and S. Brunak. Using electronic patient records to discover disease correlations and stratify patient cohorts. *PLoS Computational Biology*, 8(7):e1002141–1, 2008.

[175] S. Rosso, R. D. Angelis, L. Ciccolallo, E. Carrani, I. Soerjomataram, E. Grande, G. Zigon, and H. Brenner. Multiple tumours in survival estimates. *Eur J Cancer*, 45(6):1080–1094, 2009.

[176] L. Ryden, E. Standl, M. Bartnik, G. van den Berghe, J. Betteridge, M. de Boer, F. Cosentino, B. Jönsson, M. Laakso, K. Malmberg, S. Priori, J. Östergren, J. Tuomilehto, and I. Thrainsdottir. European society of cardiology guidelines on diabetes, pre-diabetes, and cardiovascular diseases: executive summary. *European Heart Journal*, 28:88–136, 2007.

[177] C. Salisbury, L. Johnson, S. Purdy, J. Valderas, and A. Montgomery. Epidemiology and impact of multimorbidity in primary care. *Brit J of Gen Prac*, 61(582): e12–e21, 2011.

[178] M. Salive. Multimorbidity in older adults. *Epidemiologic Reviews*, 35:75–83, 2013.

[179] I. Schäfer, E. von Leitner, G. Schön, D. koller, H. Hansen, T. Kolonko, H. Kaduszkiewicz, K. Wegscheider, G. Glaeske, and H. van den Bussche. Multimorbidity patterns in the elderly: A new approach of disease clustering identifies complex interrelations between chronic conditions. *Plos One*, 5(12):e15941, 2010.

[180] F. Schellevis. *Comorbidity of chronic diseases in general practice*. PhD thesis, University of Nijmegen, 1993.

[181] O. Schulte, G. Frigo, R. Greiner, W. Luo, and H. Khosravi. A new hybrid method for Bayesian network learning with dependency constraints. In *IEEE Symposium on Computational Intelligence and Data Mining*, pages 53–60, 2009.

[182] B. Schüz, S. Wurm, L. Warner, and C. Tesch-Römer. Health and subjective well-being in later adulthood: different health states - different needs? *Applied Psychology: Health and Well-Being*, 1(1):23–45, 2009.

[183] G. Schwarz. Estimating the dimension of a model. *Annals of Statistics*, 6(2): 461–464, 1978.

[184] S. Sciaretta, F. Palano, G. Tocci, R. Baldini, and M. Volpe. Antihypertensive treatment and development of heart failure in hypertension. *Arch Intern Med*, 171:384–394, 2011.

[185] M. Scutari. Learning Bayesian networks with the bnlearn R package. *Journal of Statistical Software*, 35(3):122, 2010.

[186] M. Scutari and R. Nagarajan. On identifying significant edges in graphical models. In A. Hommersom and P. Lucas, editors, *Proceedings of Workshop on Probabilistic Problem Solving in BioMedicine*, pages 15–27, Bled, Slovenia, 2011. Springer Verlag.

[187] M. Seltzer, W. Wong, and A. Bryk. Bayesian analysis in applications of hierarchical models: issues and methods. *Journal of Educational and Behavioral Statistics*, 21:131–167, 1996.

[188] R. Shachter. Evaluating influence diagrams. *Operations Research*, 34(6):871–882, 1986.

[189] M. Shwe, B. Middleton, D. Heckerman, M. Henrion, E. Horvitz, H. Lehmann, and G. Cooper. Probabilistic diagnosis using a reformulation of the INTERNIST-1/QMR knowledge base. I. the probabilistic model and inference algorithms. *Methods Inf Med*, 30(4):241–55, 1998.

[190] C. L. Sistrom and C. W. Garvan. Proportions, odds, and risk. *Radiology*, 230: 12–19, 2004.

[191] R. Smith. Efficient monte carlo procedures for generating points uniformly distributed over bounded regions. *Operations Research*, 32(6):1296–1308, 1984.

[192] S. Smith, H. Soubhi, M. Fortin, C. Hudon, and T. O'Dowd. Managing patients with multimorbidity; systematic review of interventions in primary care and community settings. *BMJ*, e5205:345, 2012.

[193] R. Sousa, C. Ferri, D. Acosta, E. Albanese, M. Guerra, Y. Huang, K. Jacob, A. Jotheeswaran, J. Rodriguez, G. Pichardo, M. Rodriguez, A. A. Sosa, J. Williams, T. Zuniga, and M. Prince. Contribution of chronic diseases to disability in elderly people in countries with low and middle incomes: a 10/66 dementia research group population-based survey. *Lancet*, 374:1821–30, 2009.

[194] D. Spiegelhalter. Bayesian graphical modelling: A case-study in monitoring health outcomes. *Applied Statistics*, 47(1):115–133, 1998.

[195] D. Spiegelhalter, A. Thomas, N. Best, and D. Lunn. *WinBUGS User Manual; Version 1.4*. MRC Biostatistics Unit, Cambridge, UK, 2001.

[196] D. Spiegelhalter, N. Best, B. Carlin, and A. van der Linde. Bayesian measures of model complexity and fit. *J.R. Statist. Soc. B*, 64(4):583–639, 2002.

[197] M. Studený and R. Bouckaert. On chain graph models for description of conditional independence structures. *The Annals of Statistics*, 26(4):1434–1495, 1998.

[198] M. Suojanen, S. Andreassen, and K. Olesen. A method for diagnosing multiple diseases in MUNIN. *IEEE Transactions on Biomedical Engineering*, 48(5):522–532, 2001.

[199] J. Tian, J. Pearl, and A. Paz. Finding minimal d-separators. Technical report, Computer Science Department, Cognitive Systems Laboratory, University of California, Los Angeles, USA, Februayr 1998.

[200] I. Tsamardinos, L. Brown, and C. Aliferis. The max-min hill-climbing Bayesian network structure learning algorithm. *Machine Learning*, 65(1):31–78, 2006.

[201] A. Uijen and E. van Lisdonk. Multimorbidity in primary care: Prevalence and trend over the last 20 years. *Eur J Gen Pract*, 14(1):28–32, 2008.

[202] L. Uusitalo. Advantages and challenges of bayesian networks in environmental modelling. *Ecological Modelling*, 203:312–318, 2007.

[203] V. Vaccarino, L. Badimon, R. Corti, C. de Wit, M. Dorobantu, A. Hall, A. Koller, M. Marzilli, A. Pries, and R. Bugiardini. Ischaemic heart disease in women: are there sex differences in pathophysiology and risk factors? position paper from the working group on coronary pathophysiology and microcirculation of the european society of cardiology. *Cardiovascular Research*, 90:9–17, 2011.

[204] J. Valderas, B. Starfield, B. Sibbald, C. Salisbury, and M. Roland. Defining comorbidity: Implications for understanding health and health services. *Ann Fam Med*, 7:357–363, 2009.

[205] T. van Allen and R. Greiner. Model selection criteria for learning belief nets: An empirical comparison. In *ICML2000*, pages 1047–1054, 2000.

[206] M. van den Akker, F. Buntinx, and J. Knottnerus. Comorbidity or multimorbidity; what's in a name? a review of the literature. *Eur J General Practice*, 2: 65–70, 1996.

[207] M. van den Akker, F. Buntinx, J. Metsemakers, S. Roos, and J. Knottnerus. Multimorbidity in general practice: prevalence, incidence and determinants of co-occurring chronic and recurrent diseases. *J Clin Epidemiol*, 51:367–375, 1998.

[208] M. van den Akker, F. Buntinx, S. Roos, and J. Knottnerus. Problems in determining occurrence rates of multimorbidity. *J Clin Epidemiol*, 54:675–679, 2001.

[209] L. van der Gaag, J. Bolt, W. Loeffen, and A. Elbers. Modelling patterns of evidence in Bayesian networks: A case-study in classical swine fever. In *Computational Intelligence for Knowledge-Based Systems Design*, volume 6178 of *Lecture Notes in Artificial Intelligence*, pages 675–684. Springer, 2010.

[210] J. van der Meer, A. Oerlemans, D. van Steijn, M. Lappenschaar, L. de Sonneville, J. Buitelaar, and N. Rommelse. Are autism spectrum disorder and attention-deficit/hyperactivity disorder different manifestations of one overarching disorder? cognitive and symptom evidence from a clinical and population-based sample. *Journal of the American Academy of Child and Adolescent Psychiatry*, 51 (11):1160–1172.e3, 2012.

[211] M. van der Wel, R. Jansen, J. Bakx, H. Bor, M. OldeRikkert, and C. van Weel. Non-cardiovascular co-morbidity in elderly patients with heart failure outnumbers cardiovascular co-morbidity. *European Journal of Heart Failure*, 9:709–715, 2007.

[212] F. van Dijk, M. Lappenschaar, C. Kan, R. Verkes, and J. Buitelaar. Lifespan attention deficit/hyperactivity disorder and borderline personality disorder symptoms in female patients: A latent class approach. *Psychiatry Research*, 190(2-3): 327–334, 2011.

[213] F. van Dijk, M. Lappenschaar, C. Kan, R. Verkes, and J. Buitelaar. Symptomatic overlap between attention-deficit/hyperactivity disorder and borderline personality disorder in women: the role of temperament and character traits. *Comprehensive Psychiatry*, 53(1):39–47, 2012.

[214] F. van Kouwen, S. Renooij, and P. Schot. Inference in qualitative probabilistic networks revisited. *International Journal of Approximate Reasoning*, 50(5):708–720, 2009.

[215] S. van Oostrom, H. Picavet, B. van Gelder, L. Lemmens, N. Hoeymans, R. Verheij, F. Schellevis, and C. Baan. Multimorbidity and comorbidity in the dutch population - data from general practices. *Nederlands Tijdschrift voor Geneeskunde*, 155(A3193), 2011.

[216] M. Velikova, M. Samulski, P. Lucas, and N. Karssemeijer. Improved mammographic cad performance using multi-view information: a Bayesian network framework. *Physics in Medicine and Biology*, 54:1131–1147, 2009.

[217] S. Visweswaran, D. Angus, M. Hsieh, L. Weissfeld, D. Yealy, and G. Cooper. Learning patient-specific predictive models from clinical data. *J Biom Inform*, 43:669–85, 2010.

[218] E. Vittinghoff, D. Glidden, S. Shiboski, and C. McCulloch. *Regression Methods in Biostatistics: linear, logistic, survival and repeated measures models*. Springer, New York, USA, 2005.

[219] E. K. Wei, K. Y. Wolin, and G. A. Colditz. Time course of risk factors in cancer etiology and progression. *Journal of Clinical Oncology*, 28(26):4052–4057, 2010.

[220] D. Weiner, H. Tighiouart, M. Amin, P. Stark, B. MacLeod, J. Griffith, D. Salem, A. Levey, and M. Sarnak. Chronic kidney disease as a risk factor for cardiovascular disease and all-cause mortality: A pooled analysis of community-based studies. *J Am Soc Nephrol*, 15:1307–1315, 2004.

[221] M. Wellman. Fundamental concepts of qualitative probabilistic networks. *Artificial Intelligence*, 40:257–303, 1990.

[222] M. Wellman and M. Henrion. Explaining 'explaining away'. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 15(3):287–292, 1993.

[223] T. Wiersma, Y. Smulders, C. Stehouwer, and et al. *Multidisciplinary Guideline on Cardiovascular Risk Management*. Bohn Stafleu van Loghum, Houten, The Netherlands, 2011.

[224] J. Wolff, B. Starfield, and G. Anderson. Prevalence, expenditures, and complications of multiple chronic conditions in the elderly. *Arch Intern Med*, 162: 2269–2276, 2002.

[225] A. Wong, H. Boshuizen, F. Schellevis, G. Kommer, and J. Polder. Longitudinal administrative data can be used to examine multimorbidity, provided false discoveries are controlled for. *J of Clin Epidemiol*, 64:1109–1117, 2011.

[226] S. Yang and K. Chang. Comparison of score metrics for Bayesian network learning. *Systems, Man and Cybernetics, Part A: Systems and Humans, IEEE Transactions on*, 32(3):419–428, 2002.

SUMMARY

The epidemiology of multiple chronic diseases present at the same time is referred to as *comorbidity* or *multimorbidity*. With the ageing of people multimorbidity becomes the rule rather than the exception, especially for the elderly. The human body is a complex adaptive system and very often we only see a few symptoms as a tip of the iceberg. Current statistical methodologies are not entirely suitable to analyse this phenomena as they often consider only one (primary) disease. In this thesis we have explored the usefulness of probabilistic network models in the field of multimorbidity. First we asked ourselves the question how interactions between diseases, frequently present with multimorbidity, can be best described. These interactions are often stochastic by nature and it turns out that many of the interactions can be expressed very well by using probabilistic networks, e.g., Bayesian networks. An important achievement of our research is that learning the structure of a network from data can significantly contribute to unravelling the intricate interactions that are hidden in clinical data. Another problem we faced in this research is the fact that much of the clinical data comes from multiple sources, e.g., from multiple general practices that use different kinds of electronic health care systems. This introduces a certain bias, and to be able to deal with such data we introduced a new concept called multilevel Bayesian networks. These networks can deal with any big dataset that is hierarchically structured. We applied them by investigating the simultaneous progression of chronic cardiovascular conditions, correcting for both patient and practice-related variables. Because of the network structure the progression is easier to understand. For example, it turned out that in the presence of hypertension, the observed cumulative incidence rates of combinations of cardiovascular disorders, i.e., multimorbidity, differ significantly from the expected rates. Another aspect is that in many real-life systems, interactions often participate in feedback loops. Here we adopted a qualitative viewpoint to model and understand such feedback loops. Although qualitative reasoning has its limitations, we showed that without knowing exact probabilities, we are still able to draw qualitative conclusions of the dynamics that exist in a system. The ideas in this thesis are certainly generalizable to other areas of scientific research. As an example we briefly discussed a simplified model of the Arctic summer sea-ice decline and its regional effects on the polar bear populations.

SAMENVATTING

De epidemiologie van meerdere chronische aandoeningen tegelijkertijd bij een patiënt wordt meestal aangeduid als *comorbiditeit* of *multimorbiditeit*. Met het toenemen van de gemiddelde leeftijd van de mens is multimorbiditeit meer regel dan uitzondering, vooral bij ouderen. Het menselijk lichaam is een complex adaptief systeem en vaak zien we alleen maar enkele symptomen als topje van de ijsberg. De huidige statistische methoden zijn meestal niet toereikend genoeg om dit fenomeen te bestuderen omdat ze vaak uitgaan van één (hoofd)aandoening. In dit proefschrift hebben we onderzocht of bepaalde kansmodellen met een netwerkstructuur ons meer inzicht kunnen verschaffen in multimorbiditeit. In eerste plaats vroegen we ons af hoe de interacties tussen aandoeningen, die men vaak tegenkomt bij multimorbiditeit, modelmatig het best beschreven kunnen worden. Omdat de interacties vaak stochastisch van aard zijn blijkt dat ze goed beschreven kunnen worden met behulp van een probabilistische netwerk zoals een Bayesiaanse netwerk. Een belangrijk resultaat van het onderzoek is dat het leren van de netwerkstructuur ons op het spoor brengt van de juiste intrinsieke interacties die er bestaan tussen bepaalde aandoeningen. Een ander probleem is dat veel van de patiëntgegevens betreffende multimorbiditeit vaak uit meerdere bronnen komt, zoals de elektronische patiëntgegevens van huisartsenpraktijken die ook nog eens verschillende systemen gebruiken. Als we dit feit negeren introduceren we een systematische fout in de analyses en resultaten. Om hiervoor te kunnen corrigeren introduceerden we een nieuw concept: een multilevel Bayesiaans netwerk. Deze netwerken kunnen goed omgaan met grote datasets die hiërarchisch georganiseerd zijn. Dit concept hebben we toegepast om de simultane progressie van chronische cardiovasculaire aandoeningen, daarbij corrigerend voor patiënt en huisartspraktijk gerelateerde variabelen, te analyseren. Door de samenhang van de aandoeningen in een netwerk is de progressie van multimorbiditeit beter te begrijpen. Het blijkt bijvoorbeeld dat bij de aanwezigheid van een hoge bloeddruk, de cumulatieve incidentie van combinaties van cardiovasculaire aandoeningen, en dus multimorbiditeit, veel sneller stijgt met de tijd dan verwacht mag worden op basis van de individuele incidentie. Een ander aspect van interacties is dat deze vaak participeren in een fysiologisch regulatiemechanisme. Omdat er dan vaak sprake is van feedback, vereist dit soort interacties een andere aanpak. Hier gebruiken we een kwalitatieve benadering en het gaat ons dan in het bijzonder om de aard van de interacties, dat wil zeggen, is deze positief of negatief, en is er sprake van synergie. Hoewel een kwalitatieve benadering zijn beperkingen heeft, laten we zien dat de hier gebruikte methodologie bruikbaar is om klinische conclusies te trekken, ook in geval van multimorbiditeit. De ideeën in dit proefschrift kunnen makkelijk vertaald worden naar andere domeinen. Als voorbeeld geven we de jaarlijkse afname van het ijsoppervlak op de Noordpool en de regionale effecten daarvan op ijsbeerpopulaties.

DANKWOORD

Het heeft even geduurd maar dan heb je ook wat. Vier jaar werk na een aanloop van veertig jaar. Voor hun tomeloze geduld, steun en bijstand dank ik daarom in de eerste plaats mijn ouders, broer, zus en partners. Bertus en Annie, jullie ondernemerschap is een blijvende inspiratiebron. Bert-Jan en Tanja, tevens bedankt dat jullie paranimf wilden zijn. Verder gaat mijn dank uit naar al mijn andere familieleden: ooms, tantes, neven, nichten, neefjes, nichtjes en alle partners.

Velen hebben bijgedragen aan dit proefschrift, waarvoor mijn uitdrukkelijke dank. Peter en Arjen, zonder jullie zou dit proefschrift er niet geweest zijn. Jullie ideeën en kritische noten hebben het proefschrift gevormd en gekleurd. De samenwerking was altijd prettig en ik heb veel van jullie geleerd. Joep, bedankt voor de medische feedback en het meedenken over het klinisch belang. Francois, Joke en Stefan, bedankt dat jullie de data beschikbaar hebben gesteld en voor het meedenken bij de juiste vraagstellingen. Allen bedankt voor jullie hulp bij het schrijven van dit proefschrift.

Graag bedank ik ook al mijn collega's bij ICIS, voor de samenwerking, collegialiteit, of gewoon voor de gezelligheid bij de vrijdagmiddagborrel. In een PhD kamer is het komen en gaan, en ik wil niemand tekort doen, maar in bijzonder bedank ik toch mijn kamergenoten Bas en Maarten die gedurende het gehele promotietraject mijn versie van 'Haribo macht Kinder froh' hebben moeten tolereren.

Dank ook aan al mijn vrienden en kennissen in de Achterhoek, Nijmegen, Rotterdam, Eindhoven, en weet ik wat voor oorden waar iedereen wortel heeft geschoten. Dank voor al jullie steun en toevertrouwen op zijn tijd.

CURRICULUM VITAE

Gerald Anne Martijn Lappenschaar

1970 Born in Borculo, The Netherlands.

1982-1988 Pre-university Secondary Education,
De Bouwmeester, Haaksbergen.

1988-1994 Applied Mathematics,
University of Twente, Enschede.

1995-1997 Software Engineer,
Accell Group, Heerenveen.

1997-2003 Software Engineer,
Baan Development BV, Barneveld.

2003-2010 Medicine,
Radboud University Medical Centre Nijmegen.

2004-2008 Datamanager,
Department of Psychiatry,
Radboud University Medical Centre Nijmegen.

2010-2014 PhD-candidate,
Model Based System Development,
Institute for Computing and Information Sciences,
Radboud University Nijmegen.

2014 Physician,
UWV (Employee Insurance Agency), Hengelo.