A TEMPORAL PROGNOSTIC MODEL BASED ON DYNAMIC BAYESIAN NETWORKS: MINING MEDICAL INSURANCE DATA

By Sarah Kerubo Mbaka

Supervised by Mzabalazo Ngwenya

Dissertation submitted in fulfilment of the requirements for the degree of

Master of Science in Data Science

In the Department of Statistical Sciences

Faculty of Science

University of Cape Town

Table of Contents

Acronyms

BN	Bayesian network
CPT	Conditional probability tables
DAG	Directed acyclic graphs
DBN	Dynamic Bayesian network
DPN	Diagnosis progression network
JPD	Joint probability distribution
MMHC	Max-min hill-climbing
MMPC	Max-min parents and children
RR	Risk ratio
TA	Temporal abstraction
TBN	Temporal Bayesian network

List of Figures

List of Tables

Basic Notation

$P(\zeta)$	global distribution
$P(\zeta_0)$	initial state distribution
$\theta_{i,j,k}$	vector set of dynamic network parameters
$X = \{X_1, \dots, X_n\}$	set of n random variables
$X_t = \{X_{1t,\dots}, X_{nt}\}$	set of n random variables at time slice t
$D = \{D_1, \dots, D_n\}$	training data
$(B_0, B_1, \dots, B_T)$	dynamic Bayesian network (DBN) with T time slices
$P(X_1, \dots, X_n)$	joint probability distribution of n random variables
$P(X_1, \dots, X_t)$	joint probability distribution for t consecutive time slices

ABSTRACT

A prognostic model is a formal combination of multiple predictors from which risk probability of a specific diagnosis can be modelled for patients. Prognostic models have become essential instruments in medicine. The models are used for prediction purposes of guiding doctors to make a smart diagnosis, patient-specific decisions or help in planning the utilization of resources for patient groups who have similar prognostic paths. Dynamic Bayesian networks theoretically provide a very expressive and flexible model to solve temporal problems in medicine. However, this involves various challenges due both to the nature of the clinical domain, and the nature of the DBN modelling and inference process itself. The challenges from the clinical domain include insufficient knowledge of temporal interactions of processes in the medical literature, the sparse nature and variability of medical data collection, and the difficulty in preparing and abstracting clinical data in a suitable format without losing valuable information in the process. Challenges about the DBN methodology and implementation include the lack of tools that allow easy modelling of temporal processes. Overcoming this challenge will help to solve various clinical temporal reasoning problems. In this thesis, we addressed these challenges while building a temporal network with explanations of the effects of predisposing factors, such as age and gender, and the progression information of all diagnoses using claims data from an insurance company in Kenya. We showed that our network could differentiate the possible probability exposure to a diagnosis given the age and gender and possible paths given a patient's history. We also presented evidence that the more patient history is provided, the better the prediction of future diagnosis.

INTRODUCTION

The medical industry can be thought of as a sequential process of information processing events from the initial collection of data, which includes the patient's history, physical examination and tests, a diagnosis is formed and the validated by further observation (Reid, Comptom, Grossman, & Fanjiang, 2005).
Comprehensive data collection and digitization of healthcare systems have created rich sources of data. Electronic health records are the largest mode of electronic medical data which is recorded at every point of service at a hospital. Although these data would be ideal for the study and modelling of disease generally, these data are not available to researchers. However, one source of these data, or at least a subset thereof, can be obtained from insurance companies. Insurance companies generate these data by recording billable interactions captured by the health care system on insured patients.

Though complex and broad, medical insurance claims data provide a unique possibility in care-related research. Claims data are considered a vital and powerful source of information that supports the decision-making process of health care stakeholders, researchers and policymakers regarding various aspects of the healthcare system. Claims data usually are a form of administrative data mostly collected by medical providers for billing and reimbursement purposes from insurance companies. The data have the benefit of following a relatively consistent method of capturing specific diagnoses, procedures, and drugs. Since every interaction between a patient and the medical system is used to generate a claim, the system creates a massive database of patient information which is standardized and formatted. Claims data can be seen as a holistic view of the patient's interactions with the health care system.

The general objective of clinical treatment is to change the clinical condition of a patient from a less healthy to a healthier state. Predicting the evolution of the clinical condition and future events is a natural part of this process. This process is of particular interest in patients' clinical conditions that change over time due to critical illnesses, chronic condition and type of treatment acquired. Prediction of future events by the human clinical expert is an uncertain process that is not well understood (Christakis & Lamont, 2000). A repeatable, formal, evidence-based model becomes highly desirable to improve understanding and accuracy and to reduce uncertainty and variability of these predictions. This model can be referred to as a prognostic model.

1.1 Background

A prognostic model is a formal combination of multiple predictors from which the risk probability of a specific diagnosis can be modelled for patients (SPRAINED study group, 2018). Prognostic models are increasingly becoming important instruments in medicine. These models use patient-specific parameters and risk factors to predict the future occurrence or reoccurrence of a specific disease or diagnosis. The models are used for prediction purposes of guiding doctors to make smart diagnosis patient-specific decisions or help in planning the utilization of resources for patient groups who have similar prognostic paths. For a patient with a given set of symptoms, the prognostic model translates the interaction variables to an estimate of the risk of experiencing a diagnosis within a specific period. In the past, such knowledge was mostly based on doctor's discretion and their experience and professional experience. Prognostic models have, therefore, been known to assist medical practitioners to make more accurate predictions based on a smart decision model with a broader knowledge base.

Prognostic models utilize the information that is patient-specific and not a disease or treatment-related information. The patient-specific information is used to calculate the chance of survival and life span of the disease. The model can monitor how the patient progresses in the life of the disease from low to high-risk (Cook, 2008).

There are two main categories of prognosis models: These include models at the patient population level and models at the individual patient level. Patient population models deal with uncovering patterns or discrepancies in cohorts of patients for a particular diagnosis, whereas individual patient models are used to formulate treatment paths that are unique to the patient. Prognostic models have improved since their inception from simple decision trees to guide therapy into deep statistical models built on large datasets.

Prognostic models are constructed using historical patient data. This analysis is commonly done by applying linear and generalized linear modelling methods. These approaches have three limitations, as highlighted by (Verduijn, 2007).

1. These models apply variable selection before inducing a model. Most of the time, this involves excluding many variables that are deemed irrelevant for the prognosis.
2. The resulting models work under the assumption a prognosis is a one-time event at a predefined time. As we will discuss later in this thesis, new information about the patient's health is recorded with time, and these models are not able to transform and update the prognosis.
3. The models have fixed roles of predictor variables and outcome variable to the attributes involved. This approach has limitation by not taking into account the dynamic nature of the health care progression where an outcome now informs what is likely to happen tomorrow.

Network-based prognosis models and modelling approaches overcome or avoid these shortcomings.

1.2 Network-Based Prognosis Models

There have been notable advances in the understanding of human diseases as a result of the network models. A graph is usually used to represent network-based disease progression models. The nodes represent events and edges represents the relationships between the events. Every node contains an already predetermined, mutually exclusive number of accepted values. These values represent the specific values under investigation, such as a diagnosis or believe. The edges show the direct dependencies among the nodes. If a directed edge connects node A to node B, then node A is known as a parent of node B, and node B is known as a child of node A. An edge connection A to B implies that the value at variable B relies on the value at the parent variable A (or that B is influenced, or caused, by A).

Many explicit probabilistic model classes have been proposed and analysed, starting with a simple path model (Vogelstein, et al., 1988). The list of extensions includes oncogenetic trees (Desper, et al., 1999), distance-based trees (Desper, et al., 2000), directed acyclic graphs (Simon, et al., 2000) oncogenetic tree mixture models (Beerenwinkel, et al., 2005) , conjunctive Bayesian networks and progression networks (Farahani & Lagergren, 2013) as well as new methods to infer probabilistic progression.

The modelling disease progression has been vastly applied in lifestyle diseases, for example, diabetes (Gao, Bihorel, DuBios, Almon, & Jusko, 2011) Parkinson's disease (Vu, Nutt, & Holford, 2012) , Alzheimer's disease (Zhou, Bohlen, Miller, & Unthank, 2008) and hypertension (Zhou, Shang, Li, Zhou, & Lu, 2012). Most studies have focused on one disease, and many researchers have attempted to model multiple diseases and their interaction with each other in one network.

(Jeong, Ko, Oh, & Han, 2017) investigated the root causes and risk factors of diseases using a network model and insurance claims data. In their network, following the standards set by the 10th revision of the International Statistical Classification of Diseases and Related Health Problems (ICD-10), the nodes represent diagnoses, and the network highlights their interaction. (Jeong, Ko, Oh, & Han, 2017) propose the assumption that a prior diagnosis could be a potential risk factor of a subsequent diagnosis. The latter allows the calculation of the relative risk (RR)

$$RR_{i\rightarrow j} = \frac{a \times N}{b \times c}$$

where a is the count of the $D_{i\rightarrow j}$ pair; b is the count of pairs having D_i as a prior diagnosis; c is the count of pairs having D_j as the subsequent diagnosis; and N is the total count of all diagnosis-diagnosis pairs.

Relative risk is the ratio of the probability of a diagnosis found in one category compared to the probability of the same diagnosis occurring in another group. The diagnoses are connected if they exhibited a temporal trend using Bonferroni-corrected Fisher's exact test. The disease progression network (DPN) was modelled with directionality in mind. The resulting DPN, which was based on claims data, including 775 diagnoses and 4,100 relationships between diagnosis pairs (2,464 unidirectional relationships, 1,335 even bidirectional relationships, and 301 lop-sided bidirectional relationships) formed by 5,736 edges. An illustration of the network is shown in Figure 1.1.

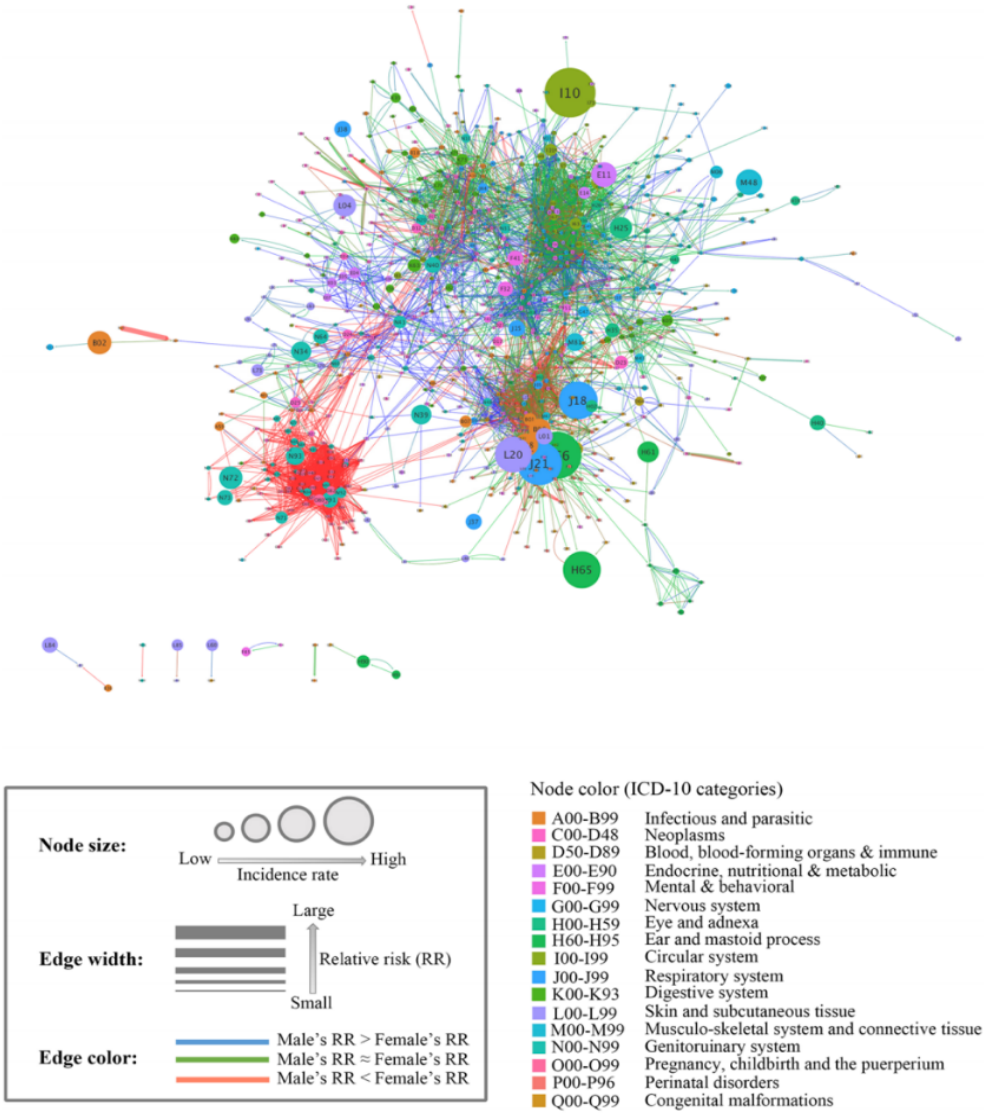


Figure 1.1 Network Disease Progression Model Using Claims Data

In the network, the size of the node is directly proportional to disease incidence. The thickness of an edge is proportional to the magnitude of the RR. According to the legend, the colours of the edges show the difference between the female RR and male RR. If we consider the existing disease networks built using clinical databases, a weighted directional network built with explanations of the risk factors such as gender and age and the diagnosis progression pairs. Although these models show the relationships between diagnoses using claims data, there are some disadvantages in doing so. A diagnosis cannot automatically be assumed to indicate a specific disease because diagnosis errors exist. However, the early prediction of disease contributes to the improvement of survival rate for patients. Therefore, reliable information concerning the progression of a disease to other diseases or from low risk to a high-risk level can be obtained, and detection time can be shortened significantly. In addition to that, the probability of misdiagnosis could be decreased. With a network that highlights the global trends that suggest which diagnosis is the next probable step in the progression with patients having different risk factors, the network here is potentially a predictive tool for the diagnoses. (Jeong, Ko, Oh, & Han, 2017) Though the network accounts for

the temporal nature of the claims data to create the DPN, time is not modelled explicitly. Dynamic Bayesian networks can potentially overcome this shortcoming.

Dynamic Bayesian networks theoretically provide a very expressive and flexible model to solve temporal problems in medicine. However, this involves various challenges due both to the nature of the clinical domain and the nature of the DBN modelling and inference process itself (Zhang, Ma, Xiao, Lin, & Yin, 2019). The challenges from the clinical domain include insufficient knowledge of temporal interactions of processes in the medical literature, the sparse nature and variability of medical data collection, and the difficulty in preparing and abstracting clinical data in a suitable format without losing valuable information in the process. Challenges about the DBN methodology and implementation include the lack of tools that allow easy modelling of temporal processes. Overcoming this challenge will help to solve various clinical temporal reasoning problems.

1.3 Relevant Works Involving DBNs in Medicine

There are several papers on the use of DBNs or use of claims data in the medical domain that aid in several clinical tasks such as diagnosis and prognosis. The models simulate medical knowledge explicitly in terms of causes and effects as inferred from the data, research from domain experts, and medical papers. They involve problems on a short timescale such as prediction of cancer progression and transplant graft survival. Some experiments involve simulated data (Zhang, Ma, Xiao, Lin, & Yin, 2019), whereas some experiments use real clinical data (Murphy, 2007). Experiments involving simulated data can be used to validate the methods, whereas experiments involving real data serve as validation and proofs of concepts of these methods' ability to answer clinical problems.

In this section, we only include works involving temporal modelling and prediction in the biomedical domain. Dynamic Bayesian networks have been applied in medicine in a minimal number of cases. Some of these studies have used simulated data, and some have used minimal amounts of real clinical data. Very few studies have used large data sets comprised of real patient data. One of the earliest works describing the use of Dynamic Bayesian networks in the biomedical domain is by (Andreassen, Hovorka, Benn, Olesen, & Carson, 1991). (Andreassen, Hovorka, Benn, Olesen, & Carson, 1991) described a combination of dynamic Bayesian networks and differential equations to model serum glucose and insulin dosing and applied it to a data set consisting of 12 patients with insulin-dependent diabetes mellitus.

Gao did a significant amount of research on dynamic models in medicine. (Gao, Bihorel, DuBios, Almon, & Jusko, 2011) utilized an ensemble of graphical models with Markov chains to find solutions in different medical domains such as neurosurgical intensive care unit monitoring ,colorectal cancer management (Date & Darwen, 2002) and palate management (McBrien, Owens, Gabbay, Niezette, & Wolper, 1990). DBNs have been used to forecast sleep apnea (Shahar, 1999), formulating a treatment path after monitoring renal failure patients who are treated using haemodialysis (Christakis & Lamont, 2000). In (Vu, Nutt, & Holford, 2012), they predicted the future progression of patients with a carcinoid tumour. The model was built by investigating and

representing the changes of various complications of the tumour, beginning with those with the most adverse effect. A comprehensive prognostic model was able to predict future states by showing the treatments and potential changes. In this model, according to the domain knowledge, an initial time, a transition interval and the end time was chosen as the model's time slices.

1.4 Aims of Research

This thesis explores the use of medical insurance claims data to model disease progression over time using DBNs for prognosis prediction of clients. The specific aims are as follows

1. Create a dynamic Bayesian model-based on medical insurance data
2. Investigate the use of the DBNs to predict patient outcomes
3. Evaluate the dynamic Bayesian network model mainly using patient-level test data
4. Illustrate the use of dynamic Bayesian networks for temporal prediction of diagnosis inpatient

1.5 Research Objectives

1. Create a disease progression model that can model multiple diseases simultaneously.
2. Create a DPN that can make inferences at the patient and population level.

1.6 Outline of Thesis

The literature review in Chapter 1 aimed to provide information to the health care system, the medical structure and discuss the issues pertinent to claims data. We then introduced prognosis models and how they have been known to assist medical practitioners to make more accurate predictions based on a smart decision model with a broader knowledge base. We explored literature on network-based prognosis models, including a case study of (Jeong, Ko, Oh, & Han, 2017) network.

Chapter 2 presents the literature on the theory of Bayesian networks and dynamic Bayesian networks. The steps used to create Bayesian networks were outlined, and these ranged from the creating of the structure, the use of the data to calculate probabilities, and how inference is performed using the results. Lastly, the use of dynamic Bayesian networks for temporal modelling was examined. The information provided in this chapter assisted in the creation of the methodology for the research.

Chapter 3 presented the modelling results and examined the different diagnosis progressions that are relevant to the research. Paths investigated include variations of population and patient-specific inference, risk factors such as age and gender and different progressions of a disease given time. Evaluation of the scenarios was done using **Bnstruct**'s inference engine, which provided information on the most influential states in the prediction of given any form of evidence. An example of a — what-if, using hypertension as a case study, the scenario was presented, and this showed the full power of the dynamic Bayesian network in modelling.

Chapter 4 presents the main findings, conclusions and the contributions of this paper.

REVIEW OF THEORETICAL FOUNDATIONS

Introduction

Bayesian networks (BNs), also known as belief networks, are graphs that belong to the family of probabilistic models. Bayesian Networks are a combination of graph theory and probability theory. It enables us to model causal and probabilistic relationships for many types of causal problems. They were introduced as a knowledge representation and inference mechanism under uncertainty using probability theory. They were successfully used in different sectors such as in medicine (Spiegelhalter, Franklin, & Bull, 1990) forecasting (Verduijn, Peek, Rosseel, & De, 2007) and speech recognition (Zweig & Russell, 1999) amongst others.

2.1 Bayesian Networks

BNs can be thought of as a database of knowledge which contains the beliefs about the interaction of variables in a system. The importance of such databases is to infer some beliefs or make some predictions or processes in the system. It is said that BNs are used to propagate beliefs throughout the network when some new information about the variables in the network is available. Consider Figure 2.1; this network investigates the interactions between diet, obesity, high cholesterol, high blood pressure and heart disease. We can infer certain beliefs; for instance, a person's diet can be used to predict the cholesterol levels and the probability of obesity. If a person is vegan, then the probability of having high cholesterol is low, which will, in turn, affect the probability of heart disease. The latter is what we call belief propagation. The value taken by one variable affects the subsequent (related) variables in the network.

The nodes on the graph represent random variables. Each node has a mutually exclusive number of values (states or beliefs). These nodes represent the variables of interest for a specific belief system such as a disease or a diagnosis. Edges represent direct dependencies (or cause-effect relationships) among variables. If a directed edge connects node A to node B, then node A is known as a parent of node B, and node B is known as a child of node A. An edge from A to B indicates that the value taken by variable B depends on the value taken by variable A (or that B is influenced, or caused, by A).
These relationships are expressed as probabilistic dependencies which are calculated through a set of conditional probability matrices. For example, if we have a variable A with states $\{a_1, a_2, \dots, a_m\}$ and conditional dependent variable B with states, $\{b_1, b_2, \dots, b_k\}$ then the conditional probabilities can be expressed with the conditional probability matrix.
The graphical representation of this distribution is a DAG with two nodes where node A is the parent node, and node B is the child node. This DAG visually specifies how the random variables depend on each other.
Formally, a Bayesian network with variables $X = \{X_1, \dots, X_n\}$ consists of:

1. The network framework that contains the probabilistic dependencies among the nodes.
2. The network parameters which are a set of local probability distributions $P(X_i|Parents(X_i))$ associated with each node. The probability distribution

indicates the effect of one node to the next. For variables that do not have any parents (i.e. the roots), their prior probability distribution is defined.

If the variables being considered are discrete, the conditional probability matrices are represented as conditional probability tables (CPTs). The CPTs define the probability or likelihood of a variable being in a particular state, given the state(s) of its parents (Baran & Jantunen, 2004). The CPT describes the effect the parent variable has on a child variable. If the data is continuous, conditional probability distributions are used. The term CPT is used in this research because it only discusses discrete variables.

Let us consider a simple model of five discrete variables: diet, high cholesterol, obesity, heart disease and high blood pressure represented by the symbols $D, O, C, BP,$ and H, respectively. Given a person's diet, they may develop obesity or high cholesterol. The probability of Obesity and high cholesterol in the general population are denoted by $P(O)$, and $P(C)$. The probability of obesity given the knowledge of the person's diet is denoted by $P(O|D)$, and the probability of high cholesterol when we know the person's diet is denoted by $P(C|D)$. The conditional probabilities of these two conditions (high blood pressure and heart disease) given the knowledge of each of these two conditions (obesity or high cholesterol) are described by the symbols $P(BP|O:C)$, and $P(D|O:C)$, respectively.

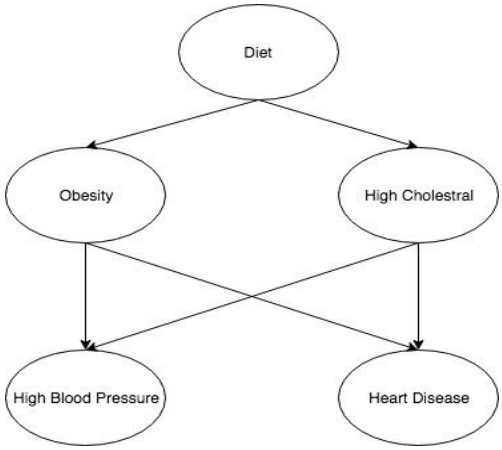


Figure 2. 1 Bayesian Network

2.1.1 D-Separation

For ease of explanation, we will move forward with the assumption that all the variables in the above network are binary, i.e., variables with only two state/cardinalities. If there were n variables in the model, then the joint probability distribution will require an order of 2^n probabilities. The latter is computationally expensive, and an optimization method is required. It should be noted that the number 2^n does not assume independence between the variables. Consider a subset of Figure 2.1 which consists of the three nodes high blood pressure (BP), high cholesterol (C), and heart disease (H). The joint probability of these three nodes is expressed as

$$P(C, BP, H) = P(C)P(C) \quad (1)$$

If we assume that high blood pressure and heart disease are conditionally independent, then equation (1) can be rewritten as

$$P(C, BP, H) = P(C)P(C)P(C) \quad (2)$$

Hence, we see that conditional independence reduces the number of terms in the joint probability distribution from $O(2^n)$ to $O(n)$. Nodes that are connected directly to each other are necessarily conditionally dependent. However, nodes that are connected indirectly may or may not be conditionally independent. The principle of dependence-separation (d-separation) provides the necessary and sufficient conditions for conditional independence in nodes that are connected indirectly. d-separation is a criterion for deciding whether a variable A is independent of another variable C, given a third variable B.

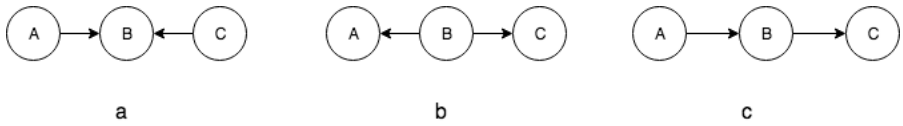


Figure 2. 2 D-Separation

Consider the three graphs in Figure 2.2. The figure shows three graphs in converging, diverging, and sequential configurations, respectively. Two nodes A and C in a graph are d-separated, if and only if there is a node B between them such that:

- the connection is sequential or diverging, and the value(state) taken by intermediate node B is known as shown in Figure 2.2 (b) and (c).
- the connection is converging, and neither the value is taken by B, nor any descendant of B is known as shown in Figure 2.2 (a).

Two nodes in a graph are conditionally independent if and only if they are d-separated.

In Figure 2.2(a), we note that variables A and C try to explain the variable B. If variable B is known, the variables A and C share the explanation for B, and hence become conditionally dependent. For example, if heart disease can be caused by both obesity and high cholesterol, if we know whether the patient has heart disease or not, the probability of obesity decreases as the probability of high cholesterol increases and *vice versa*, even if we know that obesity and high cholesterol are independent of each other. If one were to consider the relationship between these two conditions in the context of heart disease, it would lead one to think that these conditions have an inverse relationship between them.

2.1.2 Markov Blanket

A significant characteristic of Bayesian networks is that we can infer conditional dependencies between variables by visually inspecting the network's graph. To identify the nodes with which any node is conditionally independent, we need to identify its "Markov blanket". The Markov blanket of a variable is a combination of its parents, its children and its children's parents.

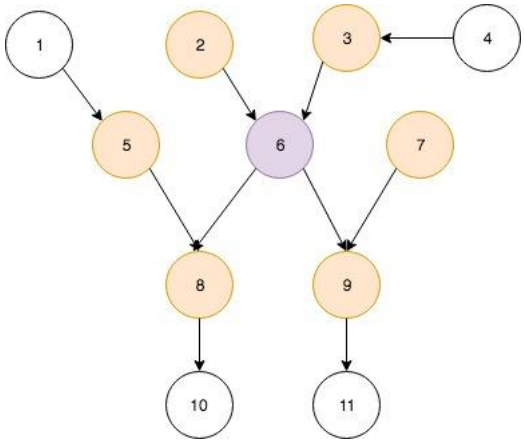


Figure 2. 3 Markov Blanket Network

In Figure 2.3, the Markov blanket of node 6 comprises the set of parents, children and its children's parents indicated by the orange nodes. The nodes in the Markov blanket include 2 and 3 as parents, 8 and 9 as children, and 5 and 7 as its children's parents of 6. The nodes 1, 4, 10 and 11 are not in the Markov blanket of 6; this implies that they are conditionally independent of node 6. This decomposition is essential when making the probability inference. Inference in the Bayesian networks setting is the task of calculating the probability of each state of a node in the network given the value of the other variables.

2.2 Dynamic Bayesian Networks

Due to the chronological format of claims data that provides the medical histories of patients, temporal abstraction (TA) of claims data aims to abstract and model claims data into meaningful higher-level temporal concepts. A variety of extensions to the Bayesian networks introduce temporality into the model. Such an extension includes dynamic Bayesian networks (DBN), which involves modelling the progress in patients' diagnosis over time using probabilistic distributions between different diagnoses, both within and across different time-slices.

A dynamic Bayesian network is a network with the repeated structure of a BN for each time slice over a specific interval (Charitos, 2001). A more appropriate term is 'temporal Bayesian networks', but the term 'dynamic Bayesian networks' has received full acceptance and more popularity. While a Bayesian network is a static model, representing the joint probability distribution at a fixed point, a DBN can represent the evolution of a system over time. In particular, DBNs allow for representing variables at multiple time points within the same network structure. Besides the static (within slice) conditional probabilistic dependencies, DBNs contain additional temporal dependencies, which are represented by edges between the time slices. Since variables in the model may be discrete or continuous, time itself may be modelled as discrete or continuous as well. A representation of a DBN with three discrete time slices is given in Figure 2.4.

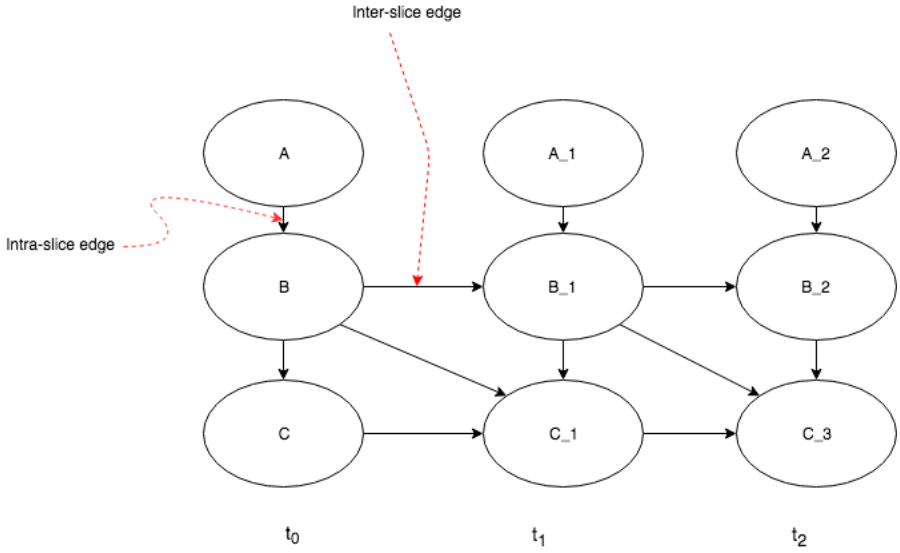


Figure 2. 4 Dynamic Bayesian Network

DBNs can be expressed as a tuple $(B_0, B_1, \dots, B_T)$ where B_0 is a Bayesian network showing the initial distribution of the variables in the first-time slice and $B_t \; for \; t > 0$ represents the temporal model for variables in time slice t and $t + 1$. DBNs represent the change of variable states at different time points where each time step is fixed. Edges represent the local or transitional dependencies among variables. Intra-slice edges represent the dependencies within the same time-slice as in a TBN. Inter-slice edges connect nodes between time slices and represent their temporal evolutions. The edges are used to highlight:

1. the evolution of a variable over time (this link is always present because the value of a variable at one time-step affects its value at the next one);
2. the relationships between different variables over time.

The network structure across the time slices does not change over time, and this characteristic is known as time invariance. Therefore, by referring to them as dynamic, we describe a dynamic system and not a network that changes structure over time. Furthermore, it is assumed that DBNs use the Markovian property: conditional probability distribution of each variable at time t, for all $t > 1$, depends only on the parents from the same time slice or from the previous time slice but not from earlier time slices.

Consider the DBN diagram of Figure 2.4
For a given time slice t, we will use $X_t = \{X_{1t,\dots}, X_{nt}\}$ to denote the n variables of the time slice and P to denote a joint probability distribution over the variables in X.

1. The transition probability for a variable i in two consecutive time slices:

$$P\left(X_{i(t-1)}\right) = P\left(parents_t(X_{it})\right) P\left(parents_{t-1}(X_{it})\right). \qquad (3)$$

where $parents_t(x_{it})$ denotes the parent set of (B_{it}) from the same time slice t, and the set $parents_{t-1}(B_{it})$ denotes the parent set of (B_{it}) from the previous time slice.

2. The joint probability distribution (JPD) for t consecutive time slices are defined by:

$$P(X_1, \dots, X_t) = P\big(parents_t(X_{it})\big) \quad (4)$$

In the example of Figure 2.4, an edge is introduced between variables A and B within the same time slice to indicate the effect of the A on the B, and between B and C at different time slices, indicating that the value is taken by the variable B at a specific time t will affect the value of variable C at time $t + 1$. As observed, no arc is extended from right to left (against the time flow). In order to comply with the direction of time, variable interactions could be perceived as causal interactions.

2.2.1 Learning with Dynamic Bayesian Networks

In most scenarios, the underlying network structure of a DBN is unknown, and the first problem is to be able to infer the structure and the parameters of the network. Bayesian networks are able to learn the parameters when given a pre-defined structure or learn a structure and the parameters at the same time. The structure and the parameters of a DBN can be learned either from data or using expert knowledge. However, it is uncommon that both the structure and parameters are learnt from the same data set. Both structure and parameter learning can be performed either as unsupervised learning, using the information provided by a data set, or as supervised learning, by interviewing experts in the fields relevant for the phenomenon being modelled. Combining both approaches is common.

There are two methods of learning and inference explored by (Murphy, 2007) they include online and offline. 'Offline' implies that all the data needed for the model is already available. Learning and inference are done with the batch of data. 'Online' implies that learning and inference are done on a rolling basis. The model is updated as soon as more data is available. Retrospective analysis of clinical data is done using offline inference, while prospective clinical decision support requires online inference.

2.2.1.1 Structure Learning

Structure learning for a dynamic network involves learning both the inter-slice and intra-slice structures. Intra-slice structure learning is similar to that with static Bayesian networks. The intra-slice connections must form a directed acyclic graph. Once intra-slice connections are learned, learning inter-slice connections becomes a variable-selection problem, where the parents of nodes in time slice t must be chosen from time slice $t - 1$. Two types of structure learning algorithms are available, namely constraint-based learning and score-based learning. Score-based learning uses predetermined scores for specific network substructures to find the structure of the complete model that maximizes the score. Constraint-based learning aims to search a model structure that satisfies a set of predefined constraints. The Max-Min Hill-Climbing (MMHC) algorithm can be referred to as a hybrid, implementing the techniques from both Constraint-based and Score-based learning (Tsamardinos, Brown, & Aliferis, 2006).

MMHC combines concepts from constraint-based and score techniques. MMHC learns the skeleton, which includes the edges with no orientation. The latter is done using an

algorithm called the Max-Min Parents and Children (MMPC). The version discussed here was proposed by (Tsamardinos, Brown, & Aliferis, 2006). Max-Min refers to the section of the algorithm that implements a heuristic approach, while the children and parent section show the output of the algorithm. MMPC is consumed by MMHC to create the structure of the BN before running a greedy search to show the orientation of the edges.

$Algorithm\ MMHC$
$Input$: $data\ (D)$;
$Output$: $Directed\ Acyclic\ Graph\ on\ the\ variables\ in\ D$
$Begin$
1. $For\ every\ node\ X\ in\ D$
$$Parents\ \&\ Children\ of\ X\ =\ MMPC(X, D)$$
$$End\ for$$
2. $greedy\ hill-climbing\ Search$
3. $Initiate\ an\ empty\ graph\ and\ perform\ Greedy\ Hill-Climbing\ with\ operators\ add-edge, reverse-edge.$
4. $Return\ the\ highest\ scoring\ DAG\ found$
5. $End\ procedure$

For every node, the algorithm identifies the set of parents and children for each variable X, after that a greedy hill-climbing search. By starting with an empty graph, the search begins by edge addition, deletion, or direction reversal. Whichever action that leads to the largest positive change in the scoring function is taken while the search continues recursively. It is important to note that a change can reduce the score. In order to terminate the algorithm, 15 consecutive changes should be observed without an increase over the maximum score ever observed during the search. The structure with the maximum score is then returned.

Structure learning need not be a fully automated process. Parts of the structure or the entire structure can be manually defined using domain knowledge, and the best structure can be chosen from these predefined models using the structure learning algorithms (Friedman, 1998). All the models in this dissertation had the structure manually defined using general knowledge and medical research. Bayesian networks in which the structure is not learnt from the data but is determined using domain knowledge are called expert systems.

2.2.1.2 Parameter Learning

The problem of parameter learning a probabilistic network is stated as follows: investigate the most probable parameters θ that explain the data. Let $D = \{D_1, D_2, \dots, D_N\}$ be the training data where $D_l = \{x_1[l], x_2[l], \dots, x_n[l]\}$ consists of instances of the Bayesian network nodes. The component θ represents the set of parameters that quantifies the network. Based on a fixed structure, parameters can be learned iteratively using expectation-maximization (EM). The EM algorithm is one of the frequently used algorithms for both parameter and structure learning in both static as well as dynamic Bayesian networks.

The expectation-maximization algorithm searches for the maximum likelihood parameter estimates in sparse data. It estimates the parameters of a model iteratively, beginning from an initial starting point. Each iteration involves an expectation (E) step and a maximization (M) step. The expectation step searches the distribution for the missing data, given the observed values for the data and the current parameter estimates. The maximization step re-calculates the parameters and returns those with the maximum likelihood, with the assumption that the distribution returned in the E step is accurate. The idea is such that each iteration gets closer to the true likelihood or remains constant (if the local maximum has been attained).
The probability distribution of all nodes can be represented as:

$$P(X) = P(X_i|parents(X_i)) \quad (5)$$

The algorithm provides each node with a conditional probability table which is denoted vector θ. The vector consists of a combination of parameters, say $P_{i,j,k}$, and is defined by:

$$P_{i,j,k} = P(X_i = X_k|parents(X_i) = X_j) \quad (6)$$

Where $i = 1 \dots n$ shows all the variables, $k = 1 \dots r_i$ describes all possible states(cardinality) taken by X_i and $j = 1 \dots q_i$ ranges all possible parent configurations of node X_i .

Algorithm Expectation Maximization EM

input: *DAG, database D, E function that calculate expectation*
output: $_{i,j,k}$

Begin
1. $t = 0$
2. *Randomly initialize the parameters*
3. *Repeat*
4. *Expectation*
Use the current parameters $P_{i,j,k}$ (t) *to estimate missing parameters*:

$$E^{(t)}(N_{i,j,k}) = \Sigma P^{(t)}(X_i = x_k \mid parents(X_i) = x_j)$$

For i from 1 *to N*

5. *Maximization*
Use estimate date to apply the learning procedure
(*for example the maximum likelihood*)

$$E_{i,j,k}{}^{(t+1)} = E(N_{i,j,k}) / E(N_{i,k})$$

6. $t = t + 1$
Until convergence ($_{i,j,k}^{(t+1)} = {}_{i,j,k}^{(t)}$)
End

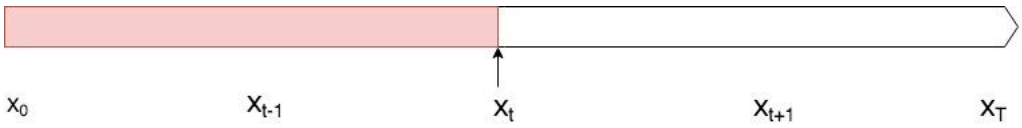
Where $N_{i,j,k}$ is the number of events in the database for which the variable X_i is in state x_k and his parents are in the configuration x_j .

2.2.2 Inference with Dynamic Bayesian Networks

After the structure and parameters of a model are determined, the model can be used to predict future events or explain past events, in a process known as inference. Inference in a DBN setting refers to the process of computing probabilities of a set of variables when given a set of known variables is provided. More specifically, having chosen a significant state within a variable, its probability across different time slices is calculated. Many of the algorithms used for inference algorithms are temporal extensions of those used for static Bayesian networks.

Assuming, $y_{0:T} = \{y_0, \ldots, y_T\}$ represents the observations up to and including time t, inference in a DBN can be done in three ways, monitoring, prediction, and smoothing (Murphy, 2007).

- Monitoring also is known as filtering, is the task of computing the current belief state for a variable X_i given all data that are available up to and including time t. To achieve this, $P(X_{it}|y_{0:t})$ will be calculated. The latter is used to investigate the current state. This concept is illustrated graphically in the figures below. The arrow denotes the time instance at which we want to perform the inference. In the shaded area the period for which we have data, and in the unshaded area the period in which data is not yet available.

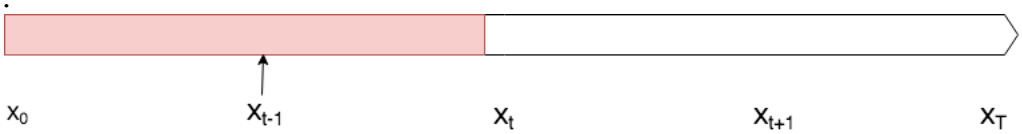


$P(X_{it}|y_{0:t})$

- Prediction, which is the task of predicting a future value at time t with all data from the previous time slices. The latter shows that the probability $P(y_{0:t})$ needs to be calculated given the states that are available of the previous time slices up to time t. Prediction is used to illustrate the effect of the available actions at time t on future time slices

$P(X_{i(t+1)}|y_{0:t})$

- Smoothing, which is the task of computing a belief state in the past at time $t-1$ given all data up to the time t (present). Thus, $P(X_{(t-1)}|y_{i:t})$. Smoothing is available to estimate the former state, as more data is available at time $t-1$ as compared to time t.



$P(X_{i(t-1)}|y_{0:t})$

METHODOLOGY: MODEL BUILDING AND EVALUATION

The dynamic Bayesian network modelling and inference methods described in Chapter 2 were applied to the claims information of patients over 2-7 years. The data was obtained from the electronic claims records of Jubilee Insurance. Before modelling and inference, the data was first transformed and abstracted into a format suitable for temporal reasoning, while minimizing the loss of information due to these transformations. After, various temporal models were constructed to reflect disease and diagnosis processes. The tools and algorithms employed for modelling and subsequent interrogation of the models are discussed in this chapter.

3.1 Pre-processing

3.1.1 Definition of Variables

Ideally, the definition of variables that are relevant to a study and their states should be carried out in a formal, well-structured process that involves relevant stakeholders or experts in the problem domain (Baran & Jantunen, 2004). However, in this study, due to practical limitation and considerations, variables were chosen based on the data provided, general knowledge and medical research.

In defining the states of the variables, either a supervised or an unsupervised approach is used. If the data is continuous, this process is called discretization. In a supervised approach, the states are set by the user after exploratory data analysis, or information from experts, policy or literature. In an unsupervised method, a suitable algorithm is used to learn and portion the data automatically; the suitable intervals for the thresholds being selected are based on statistical analysis.

3.1.2 Data Preparation

The correctness of any statistical model relies entirely upon the quality of the data fed into the model. The quality determines the precision, sensitivity and accuracy of the model. In addition to that, the technique used to infer the structure of the model as well as the parameters. The quality of training data can determine the success or failure of a model or a technique. Some machine learning models, including the DBN, cannot consume raw data extracted from the insurance systems directly. A lot of preprocessing and translation of the data need to be done. The transformation formats the data in a way that the model can consume.

The administrative claims data used for this study were managed using Oracle Database and extracted into excel files for analysis using the programming language R. Before extracting the claims data, an SQL query was used to combine the relevant variables from the respective tables. The process included selecting the IDs of patients suffering from a chronic disease, extracting their background information and the appropriate claims.

The data consisted of the following: a patient background and a claims table. Each table has various variables that were extracted. They include beneficiary ID, gender, first and second diagnosis, age, claim date and risk group. The data spans from the year 2012 to 2019. The whole sample population includes 25,597 patients whose age span from 2 to 72 years. There were about 300,000 claim incidence cases recorded. An incidence case is a claim entered on the patient's behalf due to the services rendered by the medical institution. An average of 5 – 12 incidences were recorded for each patient.

All probabilistic models can consume variables in a discrete or continuous format. They cannot consume free-text variables. The models described in this dissertation require discrete data. Claims data was discretized with a standard structure for all the variables. The total count of the variables represents the number of nodes in the network. A predetermined number of possible values is assigned to each variable. This number of possible values is referred to as its cardinality. Consider a continuous variable such as age; it can have infinite cardinality (excluding the reality of morality in humans) in order to work with it, we have to restrict it to a smaller set of values. The latter will enable us to treat it like a discrete variable. The latter is referred to as quantization.

For this data set, most variables are discrete variables, and the cardinality is predefined. Table 1 shows a sample of the variables and their first five cardinalities and the corresponding number representing the category in the model.

State	DIAGNOSIS
1	CERVICAL DISC DISORDERS
2	RECTAL BLEEDING
3	SYMPTOMATIC UTERINE FIBROIDS
4	MULTINODULAR GOITRE
5	DIABETES

State	RISK GROUP
1	BONE
2	THROAT
3	LUNGS
4	FRACTURE
5	REPRODUCTIVE

State	TREATMENT
1	ADMISSION
2	COLONOSCOPY
3	MYOMECTOMY
4	THYROIDECTOMY
5	PRESCRIPTION

Table 1 Variable cardinalities

The continuous variable included in the network is the age variable. Equal width or equal interval discretization is the most basic method among the various discretization methods. The range of the numerical values of the variable of interest was divided into the desired number of intervals or bins by dividing the range equally. The width of the intervals (i.e., the difference between the lower and upper bounds) was the same for all bins. We present the intervals chosen for the age variable using equal interval discretization in Table 2.

State Number	Age Group
1	0-10
2	11-20
3	21-30
4	31-40
5	41-60
6	Above 60

Table 2 Age Group Discretization

3.1.3 Temporal Abstraction

After the data preparation phase, the data showing the claims nodes in the model were saved in a de-normalized data table. However, the data rows for each patient may differ by days, weeks, months or even years. The latter is because patients do not visit the doctor in any regular pattern. The latter can produce a very sparse longitudinal table. This can be mitigated by temporally consolidating the data. The latter can be done by choosing one data point for each time interval represented in the model. The latter will reduce the need for performing imputation on our dataset. In our case of the insurance claims data set, the hospital visits were not in any particular interval. Hence, a good starting point was to abstract the data to select a time interval measurement such as yearly or monthly and select a diagnosis within that period as a representative of the entire time slice.

During this process, we encountered a case of both multiple records and no records observed during each chosen time-slice. The presence of missing data can be attributed to the following: there were no claims in that period; the claims were recorded and then lost. It may also mean that the client has a medical claim but did not use insurance to cater for the bill. In such cases, there are different approaches discussed to mitigate this. (Little & Rubin, 1987) refined the causes of missing data as missing at random (MAR), missing but completely at random (MCAR) and not missing at random (NMAR) (Little & Rubin, 1987). They discuss methods for dealing with missing data which apply to the clinical domain, including creating a discrete state for the clinical variable to represent missing data, or creating a separate proxy variable for each clinical variable to represent missing data.

No special treatment was implemented for missing data in this experiment. The R package **Bnstruct** (Franzin, Sambo, & Camillo, 2017) was used for model building and inference. This package implements all the algorithms we needed for model supported parameter learning with missing values by imputation. Basic imputation was performed using the k-Nearest Neighbor algorithm. The number of neighbours used was done by specifying a K value. However, in the case of multiple entries, a representative claim was chosen.

Several approaches on how to select the representative claim if there are multiple data points in a longitudinal data set are highlighted by several authors, including (Liao, 2005) . Basic approaches include calculating the mode (which is usually the mode for categorical variables) or choosing the unique measurement. In this paper, we decided to select the unique claim, because this would mean that the patient had a new diagnosis

different from the prior diagnosis. In this manner, we were able to have a temporally abstracted data with a null value in the case of missing data points and one claim in the case of multiple claims.

3.2 The Model

3.2.1 Network Construction

Creating a dynamic Bayesian model consisted of three steps: defining the nodes, defining the edges, and defining the states of the nodes. The structure of the models used in all the experiments described in this dissertation is based on domain knowledge gathered from medical literature. Structure learning algorithms are not used. The structure is based on the knowledge that chronic diseases that progress slowly are among the most common, expensive, and debilitating of all health problems and patient history are very important to be able to show the current risk areas of a patient. Therefore, the network proposed here takes the initial diagnosis and uses it to calculate the current and future state of the patient.

We first begin by modelling the nodes in the DBN. We then define the states based on the output from the discretization algorithm for the continuous variable and the categories for the categorical variables. Finally, the edges are defined. Intra-slice edges are defined first, followed by inter-slice edges

Initially, the variables contained categories encoded using characters, as shown in the previous section. The latter was changed by encoding the categories using integers in preparation of feeding it into the network. The table below contains the cardinality of the variables in the network after encoding.

Variable	cardinality
Gender	2
age	7
diagnosis	177
risk group	17

Table 3 Variable summary

The following variable naming convention was used to distinguish the variables at different time slices: each variable will contain an extension of t where t indicates the time slice. For example, the initial time slice the variable GENDER in the network was denoted as GENDER_0.

Bnstruct allows for layering of the nodes. Variables can be grouped in (numbered) layers, and a variable can only have parents in upper layers or from the same layer. In the learning step, the layering needs to be determined and fed into the model first. The first layer contains variables with no parents, and variables in layer j can have parents only in layers $i \leq j$. Nodes GENDER and AGE were encoded as layer 1 because they are risk factors. The DIAGNOSIS node is predefined as layer 2, while RISK GROUP and TREATMENT are layers 3 and 4, respectively. The layering within each time slice is

arranged such that parent nodes are those from the next layer above. However, when initializing the inter-slice edges, this rule is relaxed, and nodes are allowed to have parents from any layer from a previous time slice while having the same restriction within its time slice.

The structure of the model is presented in Figure 3.1.

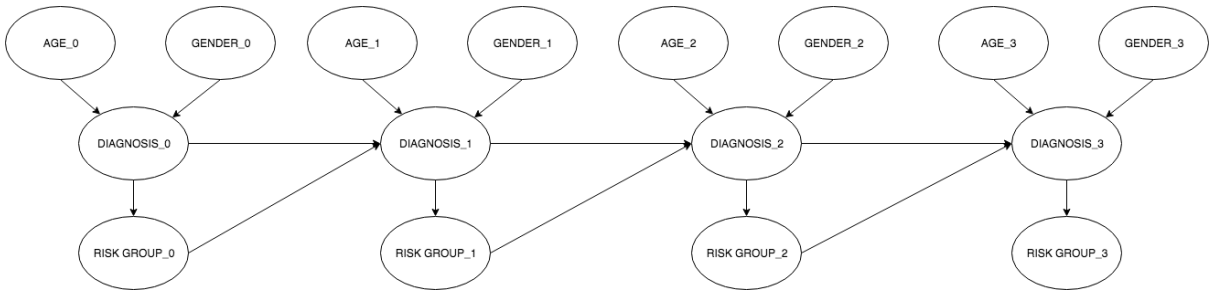


Figure 3. 1 Prognosis Dynamic Bayesian Network

The associated decomposition of the joint probability distribution that is expressed by the network is given as

$$P(\zeta) = P(\zeta_0) \cdot P(\zeta_1) \cdot P(\zeta_2) \cdot P(\zeta_3) \quad (7)$$

where

$$P(\zeta_0) = P(A_0, G_0, D_0, R_0) \ = P(A_0)\, P(G_0)\, P(A_0 : G_0) P(D_0) \quad (8)$$

$$P(\zeta_1) = P(D_0, R_0, A_1, G_1, R_1, D_1) = P(A_1)\, P(G_1)\, P(D_0 : R_0 : A_1 : G_1) P(D_1) \quad (9)$$

$$P(\zeta_2) = P(D_1, R_1, A_2, G_2, R_2, D_2) = P(A_2)\, P(G_2)\, P(D_1 : R_1 : A_2 : G_2) P(D_2) \quad (10)$$

$$P(\zeta_3) = P(D_2, R_2, A_3, G_3, R_3, D_3) = P(A_3)\, P(G_3)\, P(D_2 : R_2 : A_3, G_3) P(D_3) \quad (11)$$

(ζ_0) is the initial state distribution, $P(\zeta_1)$ the transition model, $P(\zeta_2)$ is a sensor model, and also, the transition model to $P(\zeta_2)$, which is the final sensor-observation model.

3.2.2 Network Properties

Two network structures are equivalent if the set of distributions that can be represented using one of the structures is identical to the set of distributions that can be represented using the other (Chickering, 2002). An equivalence class refers to nodes that have the same set of parameters, which means that the structure of the conditional probability tables is similar. The parameters, in this case, are said to be 'tied'. Parameter tying is a benefit of the Markov property by assuming the model to be a homogeneous Markov chain, i.e., the structure of the conditional probability tables does not change over time. If a given node in a time slice and its cohort in a previous or subsequent slice have the same set of ancestors, then they are considered to be in the same equivalence class; if not, they are then in different equivalence classes. Hence, for a model with n nodes per time slice in two time slices, the maximum number of parameters (conditional probability tables) to be learned is $2n$, which is the number of nodes in the two-time slices. However, if there are m equivalence classes in the first time slice, and $m \leq n$, then the total parameters needed to describe the model is $2n - m$, since these nodes in the first and the second time slices have the same parameters.

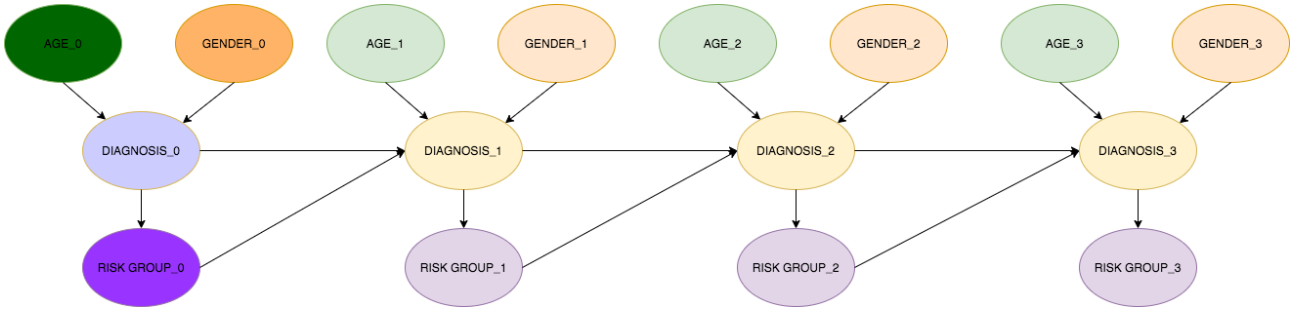


Figure 3.2 shows the model with the equivalence classes identified.

Different equivalence classes are shown using different colour schemes. At t_0 four equivalence classes are representing each node. Three nodes in the first time slice and three nodes all the subsequent time slices are in the same equivalence classes. As shown in figure 3.2, this model has only 5 equivalence classes. The distribution of DIAGNOSIS_0 is $P(A_0 : G_0)$ while the distributions at DIAGNOSIS _1, DIAGNOSIS _2 and DIAGNOSIS _3 are $P(D_0 : R_0 : A_1 : G_1) \approx P(D_1 : R_1 : A_2 : G_2) \approx P(D_2 : R_2 : A_3 : G_3)$ respectively.

During training, **Bnstruct** automatically unrolls (repeats the network structure) the DBN for all the time slices. The structure of the nodes and edges are repeated, and the parameters (conditional probability tables) are shared from the second until the last time slice. Hence, for a model with t timeslices and n nodes per time slice, instead of having tn parameters, we have $tn - (t-1)m$ parameters to learn. In addition to the reduction in complexity, parameter tying also helps to support training with an arbitrary number of time slices, and a data set where each case (patient) has data with a different number of time slices. The latter provides support for real clinical data where different patients have different lengths of claims, and hence data sets of different temporal lengths.

It is challenging to present the whole network with all of its parameters: The nodes have several states, and some CPTs consist of 4 dimensions. An overview of the nodes and model structure will, therefore, be given beforehand. As an example, the final network and the results obtained for the diagnosis of Hypertension are illustrated in Figure 3.4.

Since the risk factors for diagnosis are age and gender, these nodes can be seen in every time slice. Although we do not anticipate that the gender of a patient will change from t_0 to t_1, age is allowed to change. The age variable here was calculated from the patients' date of birth and the time of the claim. As the data provided is historical data of up to 6 years we anticipated that the patient's age would change with time and therefore we can see that the prior probabilities of the age groups change slightly from one time slice to the next. Consider the age group 5 (50-59) in Figure 3.4. The representation changes gradually.

The two nodes, AGE and GENDER, are what we characterized as layer 1, this means that they are prior distributions.

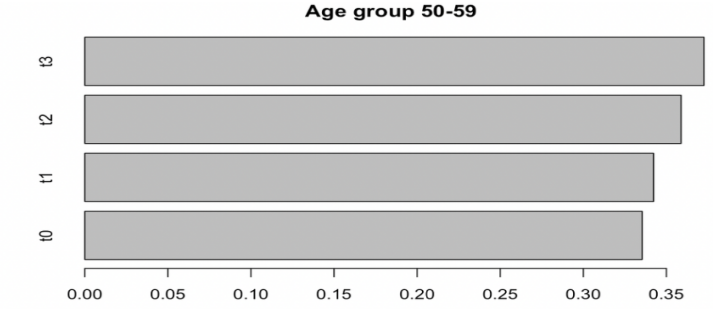


Figure 3. 2 Female Age Group 50-59 Hypertension

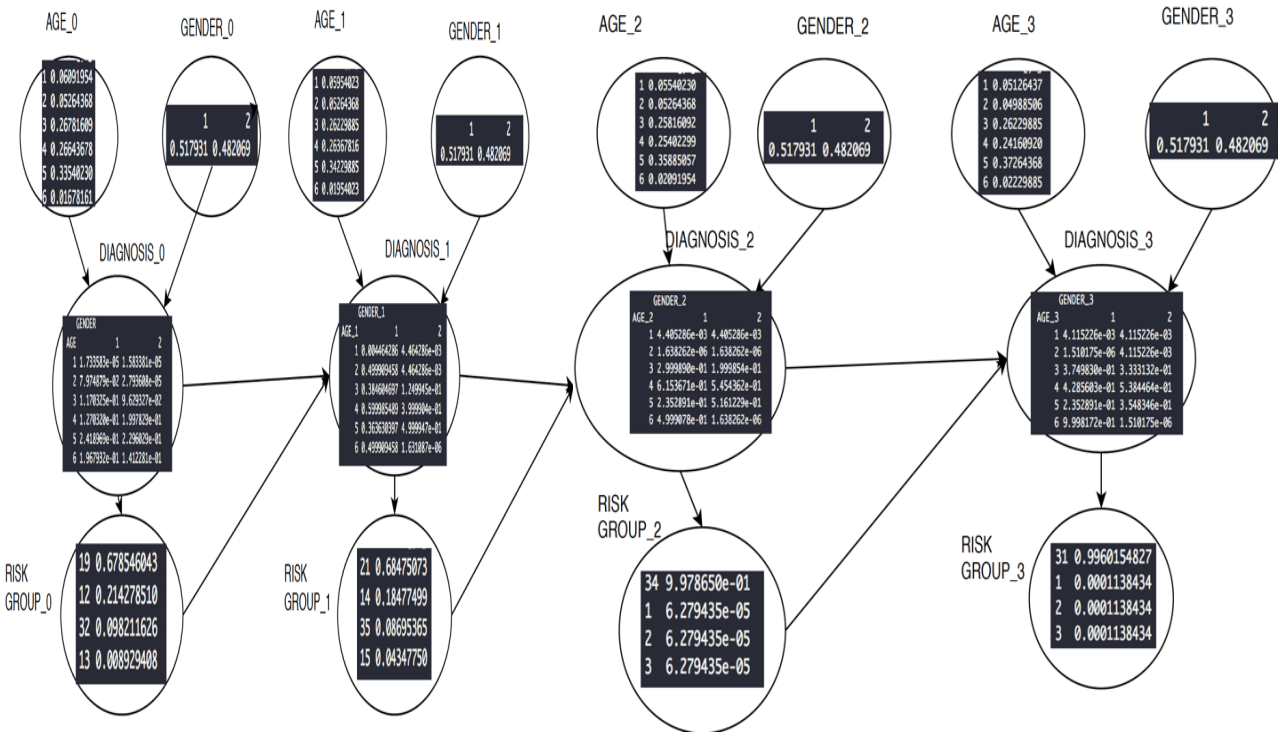


Figure 3.3 Hypertension BN Network Conditional Probability Table

At the diagnosis level, layer 2, we have an interaction of inter-slice and inter-slice edges and can be seen in the structure of the conditional probability tables. At the node DIAGNOSIS_0, this is the first diagnosis that the patient claimed. The latter is not forgetting that the patient could have been receiving treatment without claiming but for the longitudinal extrapolations of the patient's diagnosis progression, we will assume that this is the first sight of the diagnosis. At this node, we have over 200 diagnoses observed and as we can see from the model the parents to this node are Age and Gender. Therefore, the investigation done at this node is the probability of being diagnosed with a specific disease given one's risk factors.

3.3 Inference

As discussed in chapter 1, prognosis models have two applications, one is at the population level, and one is at the patient level. To demonstrate the kind and level of inference that can be conducted at a population level, we pick one diagnosis and investigate the population associated with that diagnosis. The selected diagnosis for this demonstration is hypertension. For a demonstration of inference at the patient level, the most probable diseases for various combinations of the risk factors (Age and Gender) are examined.

3.3.1 Inference: Population-Level

Patient population prognosis models focus on recognizing trends or discrepancies in groups of patients for a specific criterion. At the population level, let us look at Figure 3.5, which is associated with the diagnosis of hypertension. For both male and female, the majority of the patients with this diagnosis are between the age of 40 and 59. An important observation is that the second age group with the highest proportion in females is the Above 60 cohort, while for males it is the 31-40 cohort. From the population we have, it appears that females tend to develop hypertension later than men. The conclusions drawn above are taken from node DIAGNOSIS_0 which is the node that provides the initial distribution of patients with the various diagnosis classes given the risk factors of age and gender, $P(A_0, G_0)$.

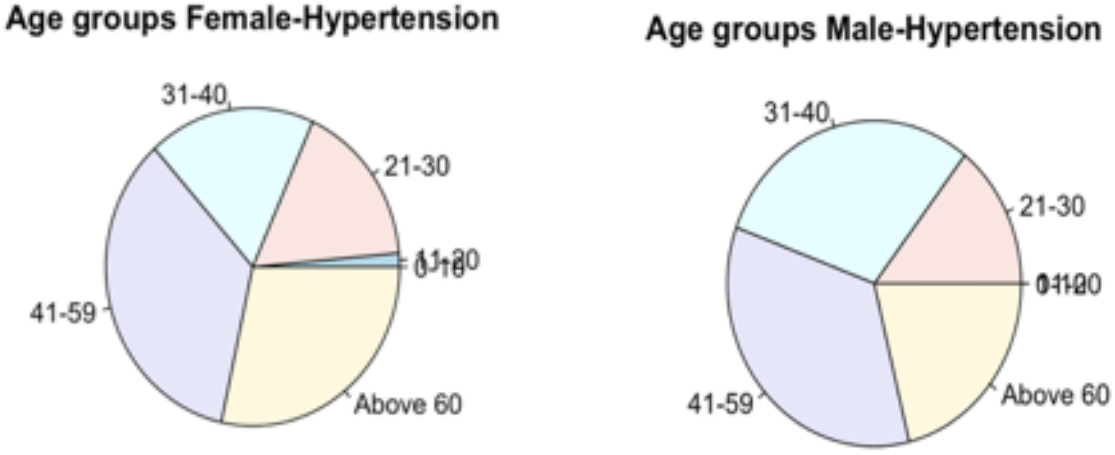


Figure 3. 4 Hypertension Age Group Distribution

Let us consider renal failure; the initial population distributions for male and female are shown in Figure 3.6. The patient distributions for hypertension and renal failure

are different. A more significant portion of the younger population seems to be affected by Renal Failure than Hypertension. Intuitively this makes sense as one would expect more adult people to suffer from high blood pressure-related diseases

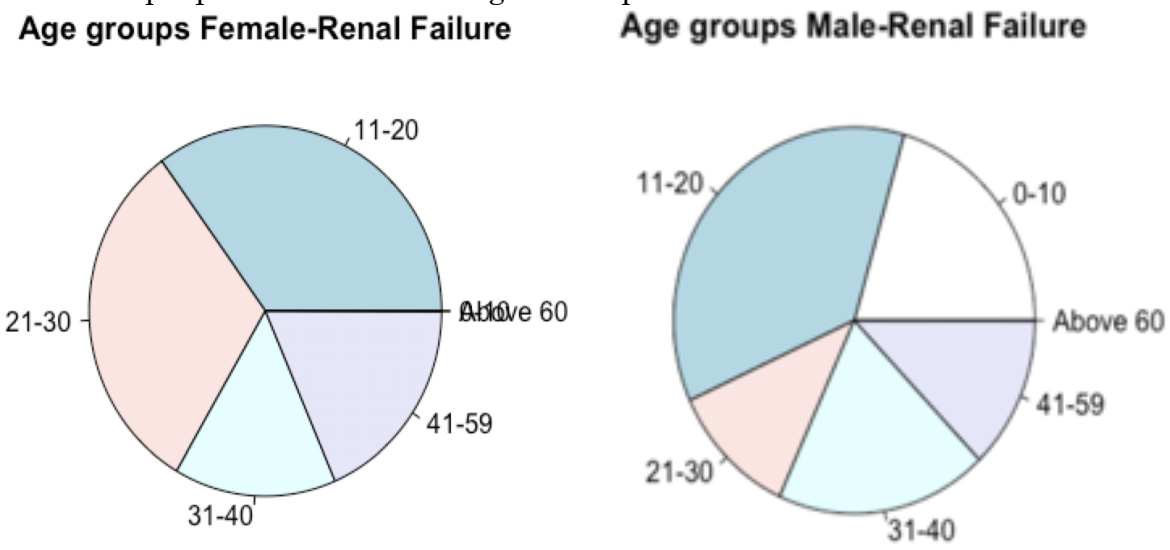
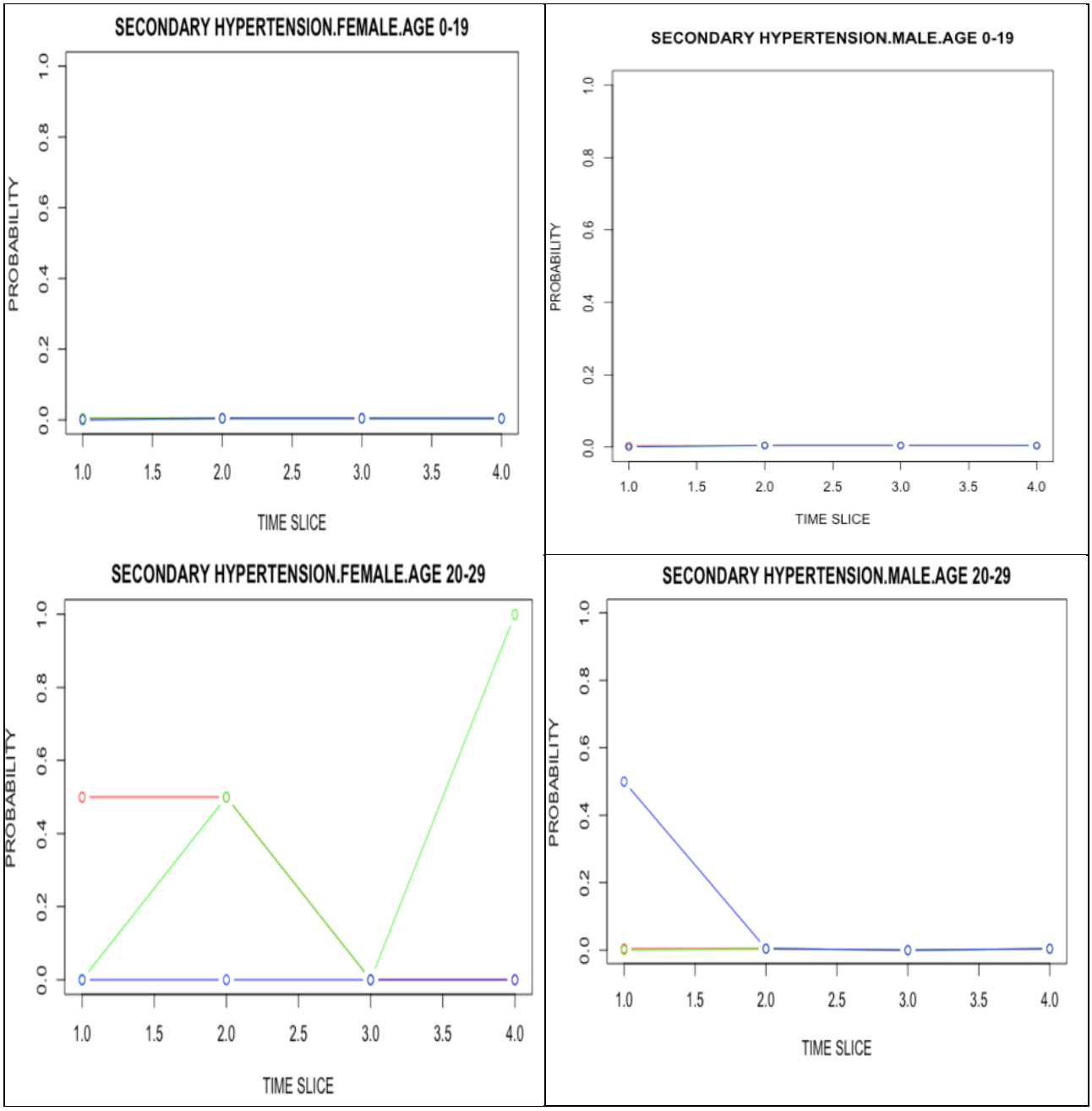


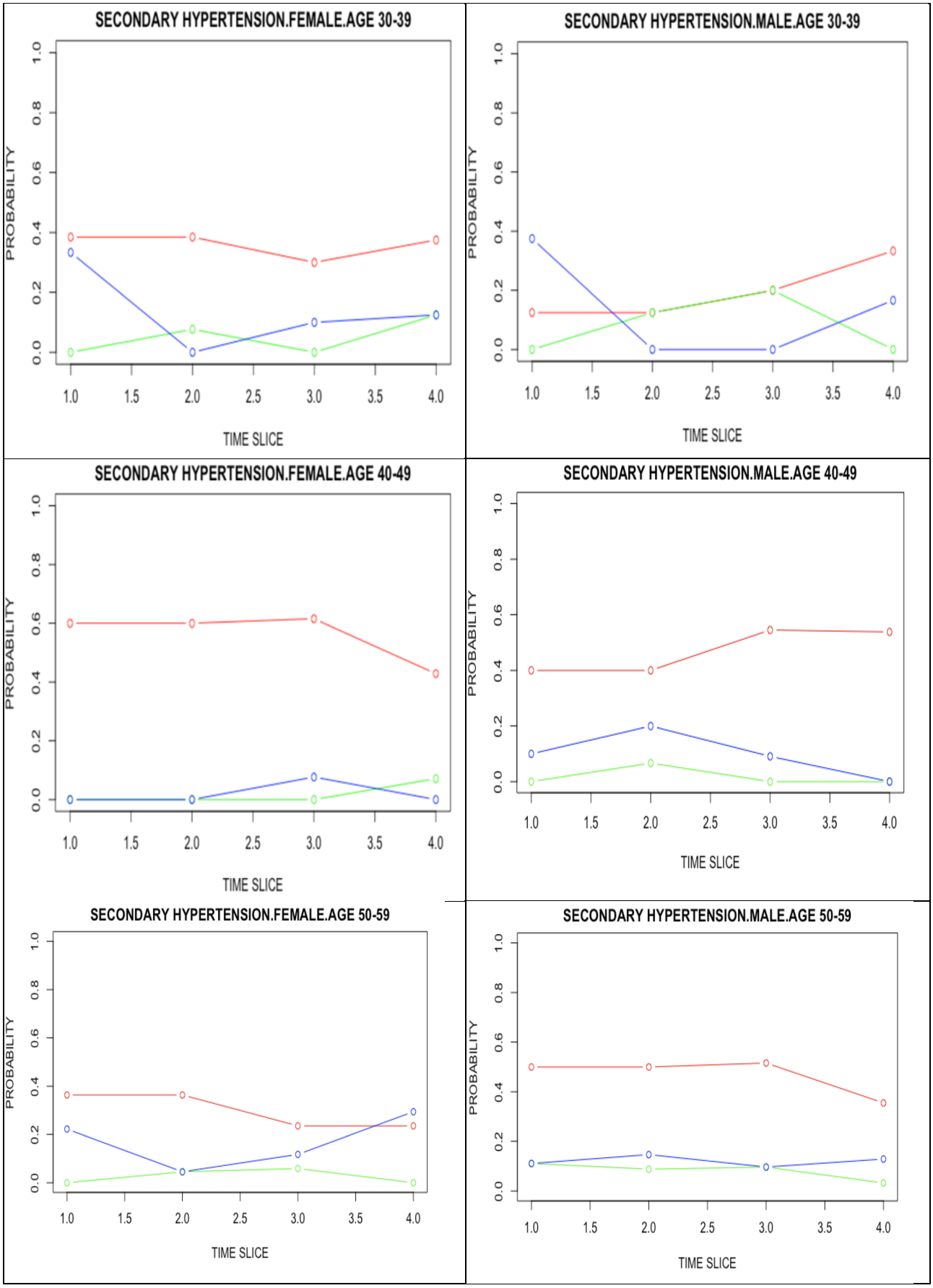
Figure 3. 5 Renal failure Age Group Distribution

The dynamic nature of the model allows for inference at the population level beyond just the initial distribution $P(\zeta_0)$. The progression of the disease over time and its interaction with other diagnoses can be inferred. We are going to look at the population of patients who were diagnosed with hypertension at any point. We demonstrate how DBNs can be used to investigate how other diagnoses interact with patients that have hypertension across the time slices.

Figure 3.7 shows the temporal extrapolation of patients who have been diagnosed with hypertension at any point in time and the interaction with diabetes and renal failure. These two diagnoses were found to be most prevalent within patients with hypertension. This observation can be verified in medical journals such as (World Health Organization, 2005) which draw a direct relationship between these three diseases.

We can observe that patients within the age of 0-19 have a very low probability of having any of the above diseases, however, from the age of 20-29, in Females, the probability of contracting hypertension increases, and it actively interacts with renal failure. In t_3, all of the patients seemed to suffer from renal failure. In male patients, the graph shows that the patients mostly have diabetes and have a low probability of hypertension or renal failure compared to their female counterparts. All the following graphs describe the interaction between the disease and how they affect the different genders and age groups. The graphs give $P(A_{i0}: G_{i0})$ at time slice 0 and $P\left(D_{i(t-1)}: R_{i(t-1)}: A_{it}: G_{it}\right)$ at the subsequent time slices


SECONDARY HYPERTENSION.FEMALE.AGE 0-19
PROBABILITY
0.0
0.2
0.4
0.6
0.8
1.0
1.0
1.5
2.0
2.5
3.0
3.5
4.0
TIME SLICE
SECONDARY HYPERTENSION.MALE.AGE 0-19
PROBABILITY
0.0
0.2
0.4
0.6
0.8
1.0
1.0
1.5
2.0
2.5
3.0
3.5
4.0
TIME SLICE
SECONDARY HYPERTENSION.FEMALE.AGE 20-29
PROBABILITY
0.0
0.2
0.4
0.6
0.8
1.0
1.0
1.5
2.0
2.5
3.0
3.5
4.0
TIME SLICE
SECONDARY HYPERTENSION.MALE.AGE 20-29
PROBABILITY
0.0
0.2
0.4
0.6
0.8
1.0
1.0
1.5
2.0
2.5
3.0
3.5
4.0
TIME SLICE


SECONDARY HYPERTENSION.FEMALE.AGE 30-39
PROBABILITY
TIME SLICE
SECONDARY HYPERTENSION.MALE.AGE 30-39
PROBABILITY
TIME SLICE
SECONDARY HYPERTENSION.FEMALE.AGE 40-49
PROBABILITY
TIME SLICE
SECONDARY HYPERTENSION.MALE.AGE 40-49
PROBABILITY
TIME SLICE
SECONDARY HYPERTENSION.FEMALE.AGE 50-59
PROBABILITY
TIME SLICE
SECONDARY HYPERTENSION.MALE.AGE 50-59
PROBABILITY
TIME SLICE

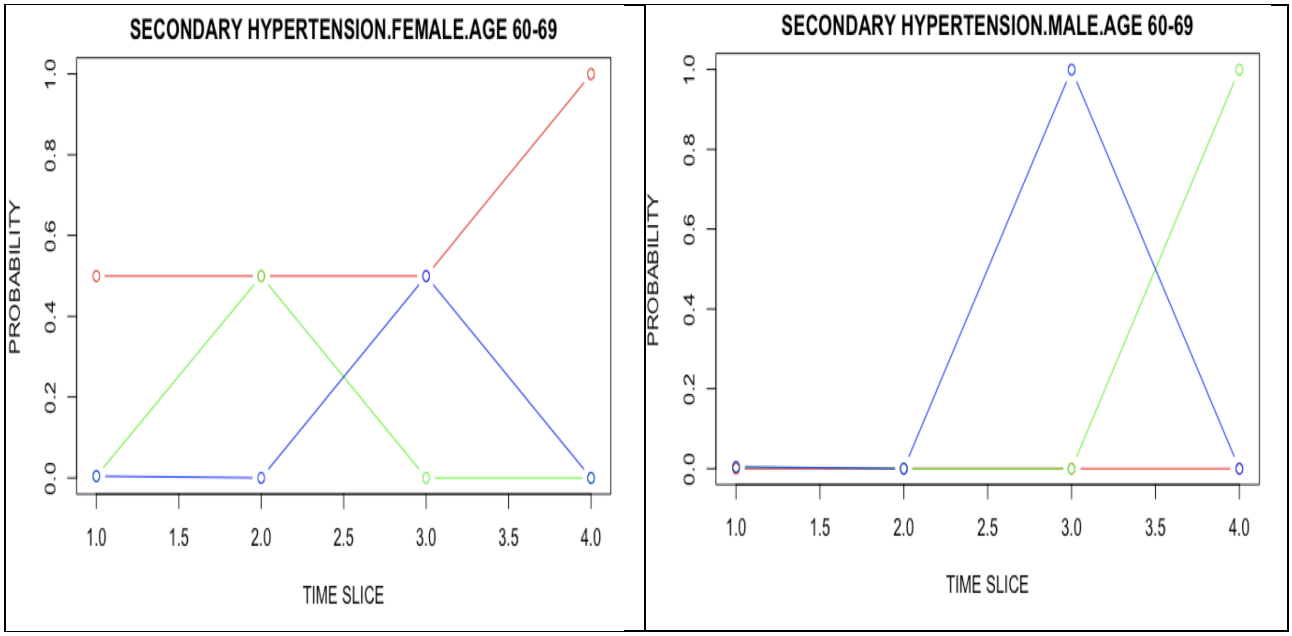


Figure 3. 6 Hypertension Time Series Analysis

3.3.2 Inference: Patient-Level

Individual patient prognosis models are used to govern treatment advice and provide patient-centred consultation. At t_0, the needed risk factors are just the age and gender of a patient, and we can be able to predict the prior probabilities of any diagnosis; $P(A_{0i}: G_{0i})$. When we consider a patient at t_1, we have more than just the risk factors, and we have previous diagnosis and risk category at t_0 , D_0 and R_0 respectively from the initial distribution which will be used to update the network for belief propagation and get a more accurate result in the subsequent time slices.

Consider a 50-year-old male patient at time slice 0. Table 4 shows the four most probable diagnosis based on the risk factors provided only; $P(A_{0(50)}: G_{0m})$.

State	Diagnosis	Probability
90	Hypertension	0.22960292
52	Diabetes	0.12155565
87	HIV	0.07428497
126	Renal failure	0.05402610

Table 4 Prior probabilities for a 50-year-old male patient

If we consider a different patient with different risk factors such as a 30-year-old female, table 5 shows the four most probable diagnosis based on the risk factors provided only; $P(A_{0(30)}: G_{0f})$. As we can note, this patient has a risk of different diseases and different probabilities from the first one. This patient-level analysis becomes better as we continue with the time slices as we will be able to know more information besides the risk factors.

State	Diagnosis	Probability
20	Breat Cancer	0.11703255
90	Hypertension	0.11703255
87	HIV	0.06301904
220	Ulcers	0.05401679

Table 5 Prior probabilities for a 30-year-old female patient

At time slice 1 we have our first interaction with the inter-slice and intra-slice edges. Node diagnosis_1 has four parents, and they include Gender_1 and Age_1 intra-slice edges meaning they belong to the same time, DIAGNOSIS_0 and Risk Group_0 as inter-slice edges which means that they are from the previous time slice. This node can be used to calculate $P(D_{0i}: R_{0i}: A_{1i}: G_{1i})$ the probability that a patient has a specific diagnosis given the risk factors at t_1 and the previous diagnosis and risk grouping from the initial state distribution, amongst other things. As a demonstration of the kind of inference that can be performed at this node, we investigated three questions:

1. What is the probability of suffering from a specific disease, diabetes, given a similar diagnosis at t_0 the first time slice?
2. The probabilities associated with the subsequent diagnosis given the diagnosis observed in the first time slice. For example, if a patient has diabetes in the first time slice, what are the most probable subsequent diagnosis at t_1,?
3. If a patient is diagnosed with diabetes in time slice 2, t_1, what did they suffer from in time slice one, t_0?

As we discussed for time slice 0, all the Conditional Probability Tables (CPTs) are decomposed to the gender and age group, therefore, to answer the questions above, we compared the different results concerning the risk factors. The questions respectively relate to the inference tasks of monitoring, prediction and smoothing.

The table below, Table 6, shows the probability of the second diagnosis being diabetes if the prior diagnosis was also diabetes. This table answers to question 1. This question is able to show how different gender and age groups deal with disease progression. Although we are tracking the same disease from t0 to t1 we investigate how different groups move within the same diagnosis. The subsequent questions 2 & 3 tackle interaction with other diagnoses

Age Gender	Female	Male
0-10	1.631087e-06	0.999636268
11-20	0.6665860	0.499909458
21-30	5.438282e-07	0.374983072
31-40	0.4999095	0.199992853
41-60	0.3333200	0.277772228
Above 60	0.0044642	0.004464286

Table 6 Probability of having diabetes in t_1 given a diabetes diagnosis in t_0

To answer question 2, table 7 below shows the most probable diagnosis in t_1 given that the patient had diabetes in t_0. The table shows the probabilities for the age group 41-

60, male and female. This analysis can be used to conclude that diabetes can cause hypertension in male within age group 41-60, 11.11% of the time while in females 22.22% of the time. The table also highlights that diabetes in $t_0$can cause coronary heart disease, blindness and low vision in male patients and asthma, renal failure in female patients and diabetes, hypertension in both.

State number	Diagnosis	Probabilities		State number	Diagnosis	Probabilities
62	Diabetes	0.27777223		62	Diabetes	0.3333200
97	Hypertension	0.11110895		97	Hypertension	0.2222134
30	Blindness and low vision	0.05555452		19	Asthma	0.1111068
42	Coronary heart disease	0.05555452		145	Renal failure	0.1111068

Table 7 Diagnosis probabilities in t_1 for a Male in Age group 5 (left) Female Age group 5 (right)

Finally, table 8 below answers question 3 showing the major causes of diabetes in t_1. The table shows the probabilities for the age group 41-60, female and male. This analysis can be used to conclude that the diagnosis of epistaxis and pneumonia in females and Retinal disorders and Hyperglycemia in males are the most likely initial diagnoses for patients who are later diagnosed with diabetes at t_1 the second time slice.

State number	Diagnosis	Probabilities		State number	Diagnosis	Probabilities
68	Epistaxis	0.9996363		175	Retinal disorders	0.9996363
183	Pneumonia	0.9996363		88	Hyperglycemia	0.999636268
52	Diabetes	0.3333200		52	Diabetes	0.277772228
206	Symptomatic uterine fibroids	0.2499776		90	Hypertension	0.147057291

Table 8 Diagnosis probabilities in t_0 for a Female in Age group 5 (left) Male Age group 5 (right) with diabetes in t_1

3.3.3 Model Evaluation

In this section, we are going to infer the diagnosis of a patient at t_1, t_2 and t_3 given various amounts of information on the prior variables. The latter should give some kind of indication of the usefulness and performance of the dynamic Bayesian models in general and the one constructed in this thesis in particular.

Table 9 below shows 5 patients as a sample of the inference. It shows the initial diagnosis and the predicted diagnosis at t_1, t_2 *and* t_3 given just the information from time slice t_0. The green diagnoses are observed from the data while the rest time slices were omitted, and the inference was made. The blue diagnoses were correctly predicted while the red ones were incorrect. The prediction accuracy for the sample of patients is

43%, 32% and 0% for the time slices t_1, t_2 *and* t_3 respectively. The latter gives an average prediction accuracy of 29%.

Gender	A	$t_0; D_{0i}\|y_{0:0}$		$t_1; D_{1i}\|y_{0:0}$		$t_2; D_{2i}\|y_{0:0}$		$t_3; D_{3i}\|y_{0:0}$	
Female	5	52	Diabetes	62	Diabetes	52	Diabetes	73	Diabetes
Female	5	52	Diabetes	62	Diabetes	52	Diabetes	73	Diabetes
Female	5	87	Hypertension	97	Hypertension	89	Hypertension	109	Hypertension
Male	4	87	Hypertension	97	Hypertension	89	Hypertension	109	Hypertension
Male	4	102	Liver failure	112	Liver failure	107	Liver failure	128	Liver failure

Table 9 Diagnosis predictions on t_1

The inference was also made with evidence from t_0 *and* t_1 and predict the outcome at t_2 and t_3 was predicted. The results of this experiment are given in table 10. The prediction accuracy for this run was 63% at both time slices.

Gender	A	$t_0; D_{0i}\|y_{0:0}$		$t_1; D_{1i}\|y_{0:1}$		$t_2; D_{2i}\|y_{0:1}$		$t_3; D_{3i}\|y_{0:1}$	
Female	5	52	Diabetes	97	Hypertension	89	Hypertension	109	Hypertension
Female	5	52	Diabetes	62	Diabetes	52	Diabetes	73	Diabetes
Female	5	87	Hypertension	92	HIV	84	HIV	106	HIV
Male	4	87	Hypertension	62	Diabetes	80	Glaucoma	73	Diabetes
Male	4	102	Liver failure	112	Liver failure	107	Liver failure	97	GERD

Table 10 Diagnosis predictions on t_2 and t_3

The inference was also made with evidence from three-time slices, and the outcome predicted at t_3. The results for the sample of patients for this run are displayed in table 11. The prediction accuracy here was 81%.

Gender	A	$t_0; D_{0i}\|y_{0:0}$		$t_1; D_{1i}\|y_{0:1}$		$t_2; D_{2i}\|y_{0:2}$		$t_3; D_{3i}\|y_{0:2}$	
Female	5	52	Diabetes	97	Hypertension	134	Migraine	109	Hypertension
Female	5	52	Diabetes	62	Diabetes	198	Retinal disorder	109	Hypertension
Female	5	87	Hypertension	92	HIV	84	HIV	106	HIV
Male	4	87	Hypertension	62	Diabetes	80	Glaucoma	73	Diabetes
Male	4	102	Liver failure	112	Liver failure	107	Liver failure	97	GERD

Table 11 Diagnosis predictions on t_3

The series of tables, 9, 10 and 11 show that with more evidence, the predictions are more accurate. It is, however, essential to note that although the prediction at any time may be wrong, the diagnosis presented here is the state with the highest probability. There may be other diagnoses with significant probabilities that could be considered for any given patient at any given time. For example, although the last patient's diagnosis at t_3 is incorrect, when inspecting the CPTs the correct diagnosis, liver failure, had the second-highest probability.

DISCUSSION OF RESULTS AND CONCLUSIONS

4.1 Discussion

Chapter 3 presented the modelling results and examined the different diagnosis progressions that are relevant to the research. Paths investigated include variations of population and patient-specific inference, risk factors such as age and gender and different progressions of a disease given time. It is scarce to find a model that can perform a wide range of inference, like the one presented in this paper. Some models focus on performing inference on a population level and others strictly at the patient level. Evaluation of the scenarios was done using **Bnstruct**'s inference engine, which provided information on the most influential states in the prediction given any form of evidence. An example of a ―what-if analysis, using hypertension as a case study, was presented and this showed the full power of the dynamic Bayesian network as a tool for building flexible prognostic models capable of a wide range of inference.

Main findings

4.1.1 Population Inference

From the network, we have demonstrated that DBNs can be used to follow diagnoses in patients with chronic diseases successfully. The model can find patterns and associations of chronic diseases and follow populations. We were able to determine the probability of any diagnosis over time for the different risk groups. For instance, we were able to highlight the population of hypertension and discover the associated diagnosis within the patients and probabilities of shifting between one to the other. This information is valuable, specifically for the medical practitioners; it can help to identify possible diagnoses and take the necessary steps such as tests to confirm the diagnosis and take the appropriate treatments. The detection time could be shortened considerably, and the likelihood of misdiagnosis could be decreased.

This information is also suitable for the insurance company for planning and enrolment on the wellness program. By planning, the company can be used to predict which chronic diseases a patient is at risk. In order to know the cost to sustain each diagnosis, it can be possible to plan or generate *special* premiums associated with the specific care needed by such patients. Besides planning financially, Jubilee insurance has a wellness program that targets patients with chronic diseases such as hypertension, diabetes and heart failure. The wellness program has targeted activities, treatments and remedies that help patients cope/live with specific diseases and sometimes prevent or cure them. Given a patient's history, we can be able to predict if a patient will likely develop a specific disease and take the necessary precautions to prevent it in accordance with the wellness program.

4.1.2 Patient-Level Inference

The model was able to predict the outcome of a diagnosis given different levels of evidence. Given the initial distribution at t_0 the model predicted the diagnosis at the different time slices with their accuracies found in table 11.

t_1	24%
t_2	34%
t_3	18%

Table 12 Diagnoses accuracy given t_0 as evidence

Given the data in t_0 and t_1 as evidence, the inference performed on the subsequent time slices yielded the accuracy found in table 12.

t_2	56%
t_3	49%

Table 13 Diagnoses accuracy given t_0 *and* t_1 as evidence

Given t_0, t_1 *and* t_2 the accuracy in predicting the diagnosis at t_3 improves to 63.45%. We can, therefore, conclude that the more evidence the model has, the better the accuracy of the predictions.

The accuracy calculated here is extracted as the diagnosis with the highest probability. For each instance, if we extract the three most probable diagnosis from the model, the accuracy of the model changes as was discussed in chapter 3.

This model may be suitable for use in the real-world clinical setting for early detection of chronic diseases if it overcomes the following limitations:

1. The diagnoses included at t_0 are not necessarily the inception of the patient's medical history and may have, therefore captured a patient's progress halfway or even in its late stages. Though the progress is also very crucial to follow, it is instrumental in having a higher percentage of the data to capture with the complete chronological order of the patients' progress.
2. In order to ensure the validity and accuracy of the model, more patients for each of the diagnosis needs to be captured and tracked over a more extended time period. As can be noted from the data, some diagnosis like neutrophilia has been recorded a handful of times. The latter leads the model to be inaccurate when it comes to inference related to that diagnosis.

4.2 Conclusions

Dynamic Bayesian networks generalize a large class of probabilistic temporal reasoning techniques that include hidden Markov models and Kalman filter models. DBNs provide a powerful formalism to perform learning and inference with models that have complex causal probabilistic relationships within and across instances of time. Pathophysiological processes and clinical practice workflow are inherently temporal processes. The complexity of medical science and the practice of medicine prompt the need for clinical decision support tools that can help with temporal modelling and prediction. Dynamic Bayesian networks are an ideal candidate for application in the medical domain to address these challenges.

Despite the success of DBNs and related techniques in other fields such as engineering, finance, economics, speech recognition, and genomic and proteomic modelling, they have

not been used to a significant extent in clinical medicine. Challenges to their adoption include the difficulty in modelling clinical processes using a temporal model, creating the model structure, data aggregation, consolidation and discretization, support for variable-length temporal processes, learning and inference with missing data, and ease of data binding for learning and inference. These are the challenges that motivated this research. This research addresses most of these challenges, as described in this dissertation.

4.2.1 Contributions of Research

The research contributed to and explored various aspects of using Dynamic Bayesian models as prognostic models. The various technical contributions to these fields are as follows.

Model Structure:
We explored and described the structure of the model with discrete variables including the nodes, edges, and the states of the model. The nodes and edges can often be discovered from medical literature or by interviewing clinicians. We described with evidence the need to avoid conditional independence and d-separation issues and how to solve these issues.

Temporal Data Aggregation, Consolidation, and Abstraction:
Temporal data aggregation and consolidation techniques were also explored. We described data preparation techniques that can be generalized to temporal reasoning problems in medicine. We described methods that can be used to aggregate and consolidate clinical temporal data from many different data sources into a uniform denormalized relational database table. We then described a method to perform temporal abstraction to select a representative data point for each time interval.

Diagnosis Progression Modelling Using Claims Data:
A temporal network within built explanations of the effects of risk factors such as gender and age and previous diagnoses was explored. We showed that our network could differentiate the possible probability of exposure to a diagnosis given the age and gender and possible paths given a patient's history. We also presented evidence that the more patient history is provided, the better the prediction of future diagnosis. We were able to conclude that if reliable data on the disease progression to other diagnoses within each age group and gender was captured, the diagnosis time of the model could be improved, and the likelihood of misdiagnosis could be avoided. The DBN shows the disease progression at a population level and could suggest the most appropriate next step in the treatment process; the model here can be used as a prescriptive tool. Lastly, the aim is to develop a clinical-friendly model by taking into account other treatment streams such as drugs and treatment methods applied.

4.2.2 Limitations

Several limitations of the methods and tools described in this dissertation have been identified. The following sections discuss these limitations.

1. The models and tools only support discrete nodes at present, due to a limitation of

the learning and inference algorithms that are used.

2. The algorithms also do not support continuous-time DBNs or DBN models

with time slices of variable width.

4.2.3 Future Research

The extent to which DBNs can fulfil the objectives depends on the availability of accurate and complete datasets. In order to improve confidence in the network, more data is, in terms of several patients and length of records, needed to improve the confidence of the diagnosis path and discovery of new paths. Moreover, there is a need to include more variables besides the diagnosis. These variables include treatment, drugs and test results. The latter could improve the specificity of inference and diagnosis paths.

The use of expert knowledge has been shown to be vital in various aspects of modelling, firstly in developing the structure of the network and also in estimating probabilities when data is incomplete, or some immeasurable variable needs to be included. Although automatic structure mining techniques are useful for defining relationships, these techniques cannot be used solely, without expert knowledge, especially with limited data. In future, the active involvement of the medical profession should be leveraged in creating the network structure and the data manipulation.

We intend to build the model and tune it for online learning and online inferencing to allow the model to capture more data and be more accurate.

Bibliography

Andreassen, S., Hovorka, R., Benn, R., Olesen, K., & Carson, E. (1991). A model-based approach to insulin adjustment. *Artificial Intelligence in Medicine*, (pp. 239–248).

Baran, E., & Jantunen, T. (2004). Stakeholder consultation for Bayesian decision support systems in environmental management. *Proceedings of the Regional Conference on Ecological and Environmental Modelling*, 15-16.

Beerenwinkel, N., Däumer, M., Sing, T., Rahnenführer, J., Lengauer, T., Selbig, J., & Kaiser, R. (2005). Estimating HIV evolutionary pathways and the genetic barrier to drug resistance. *The Journal of Infectious Diseases, 1953–1960.*

Charitos, T. (2001). *Reasoning with Dynamic Networks in Practise.* Utrecht University.

Chickering, D. M. (2002). Learning Equivalence Classes of Bayesian-Network Structures. *Journal of Machine Learning Research*, 445-498.

Christakis, N., & Lamont, E. (2000). *Extent and determinants of error in doctor's prognoses in terminally ill patients.* BMJ.

Cook, N. (2008). Statistical evaluation of prognostic versus diagnostic models: beyond the ROC curve. *Clin Chem*, 17-23.

Cook, N. (2008). Statistical evaluation of prognostic versus diagnostic models: beyond the ROC curve.

Cook, N. (2008). Statistical evaluation of prognostic versus diagnostic models: beyond the ROC curve. *Clin Chem*, 17-23.

Date, C., & Darwen, H. (2002). *Temporal Data and the Relational Model.* San Francisco: Morgan Kaufmann Publishers Inc.

Desper, R., Jiang, F., Kallioniemi, O., Moch, H., Papadimitriou, C., & Schäffer, A. (1999). Inferring tree models for oncogenesis from comparative genome hybridization data. *Journal of Computational Biology.*

Desper, R., Jiang, F., Kallioniemi, O., Moch, H., Papadimitriou, C., & Schäffer, A. (2000). Distance-based reconstruction of tree models for oncogenesis. *Journal of Computational Biology*, 789–803.

Farahani, S., & Lagergren, J. (2013). *Leaning oncogenetic networks by reducing to mixed integer linear programming*. PLoS ONE.

Franzin, A., Sambo, F., & Camillo, B. (2017). *Bnstruct: an R package for Bayesian Network structure learning in the presence of missing data.Bioinformatics, (pp. 1250-1252).* Oxford University Press.

Friedman, N. (1998). *The Bayesian structural EM algorithm. (pp. 129 - 138).* University of California, Berkeley.

Gao, W., Bihorel, S., DuBios, D. C., Almon, R. R., & Jusko, W. J. (2011). Mechanism-based disease progression modeling of type 2 diabetes in Goto-Kakizaki rats. *Journal of Pharmacokinetics*.

group, S. s. (2018). *Prognostic models for identifying risk of poor outcome.*

group, S. s. (2018). *Prognostic models for identifying risk of poor outcome.* Health Technology Assesment.

Holford, N. (2012). Clinical pharmacology = disease progression + drug action. *British Journal of Clinical Pharmacology.*

Hope, L., & Korb, K. (2004). *A Bayesian Metric for Evaluating Machine Learning Algorithms.*

Jeong, E., Ko, K., Oh, S., & Han, H. W. (2017). Network-based analysis of diagnosis progression patterns using claims data. *Scientic Reports*.

Liao, W. (2005). Clustering of time series data–a survey. *Pattern Recognition.*, (pp. 1857–1874).

McBrien, P., Owens, R., Gabbay, D., Niezette, M., & Wolper, P. (1990). TEMPORA: a temporal database transaction system. *I. IEE Colloquium on Temporal Reasoning.*

Murphy, K. (2007). *Dynamic bayesian networks: representation, inference and learning.* U.C. Berkeley.

Reid, P., Comptom, W., Grossman, J., & Fanjiang, G. (2005). Building a Better Delivery System: a New Engineering/Health Care Partnership. *National Academies Press*, 23-24. Retrieved from https://pubmed.ncbi.nlm.nih.gov/20669457/

Shahar, Y. (1999). Timing is everything: temporal reasoning and temporal data maintenance in medicine. *Medicine, Lecture Notes In Artificial Intelligence,*, 30 -46.

Simon, R., D, R., Alberts, D., Taetle, R., Trent, J., & Schäffer, A. (2000). Chromosome abnormalities in ovarian adenocarcinoma: III. using breakpoint data to infer and test mathematical models for oncogenesis. *Genes Chromosomes Cancer*, 106–120.

Spiegelhalter, D., Franklin, R., & Bull, K. (1990). Assessment, criticism and improvement of imprecise subjective probabilities for a medical expert system. *5th Annual Conference on Uncertainty in Artificial Intelligence*, (pp. 294 -295). UAI'89.

SPRAINED study group. (2018). Prognostic models for identifying risk of poor outcome. Health Technology Assesment.

Suchániekа, P., Mareckib, F., & Bucki, R. (2014). Self-learning bayesian networks in diagnosis. *18th International Conference on Knowledge-Based and IntelligentInformation & Engineering Systems*, (pp. 1426 – 1435) Science Direct.

Tsamardinos, I., Brown, L., & Aliferis, C. (2006). The max-min hill-climbing Bayesian network structure learning algorithm. *Machine Learning*, 31-78.

Verduijn, M. (2007). Prognostic Methods in Cardiac Surgery and Postoperative Intensive Care. *PhD Thesis, Eindhoven University of Technology.*

Verduijn, M., Peek, N., Rosseel, P., & De, J. (2007). Prognostic Bayesian networks. Rationale, learning procedure, and clinical use. *Journal of Biomedical Informatics*, (pp. 609-618).

Vogelstein, B., Fearon, E. R., Hamilton, S. R., Kern, S. E., Preisinger, A. C., Leppert, M., & Bos, J. L. (1988). Genetic alterations during colorectal-tumor development. *New England Journal Medicine*, 525-532.

Vu, T. C., Nutt, J. G., & Holford, N. H. (2012). Progression of motor and nonmotor features of Parkinson's disease and their response to treatment. *Britain Journal of Clinical Pharmacology*, 267–283.

World Health Organization. (2005). Clinical guidelines for the management of hypertension. *Emro Technical Publications.*

Zhang, N. (1998). Probabilistic Inference in Influence Diagrams. *Computational Intelligence*, 475 - 497.

Zhang, T., Ma, Y., Xiao, X., Lin, Y., & Yin, F. (2019). Dynamic Bayesian network in infectious diseases surveillance: a simulation study. *Scientific Reports*, 10376.

Zhou, X., Bohlen, H. G., Miller, S. J., & Unthank, J. L. (2008). American Journal of Physiology-Heart and Circulatory Physiology. *NAD(P)H oxidase-derived peroxide mediates elevated basal and impaired flow-induced NO production in SHR mesenteric arteries in vivo*, 295.

Zhou, X., Shang, D., Li, L., Zhou, T., & Lu, W. (2012). Modeling of angiotensin II–angiotensin-(1-7) counterbalance in disease progression in spontaneously hypertensive rats treated with/without perindopril. *Pharmacological Research*, 177-184.

Zweig, G., & Russell, S. (1999). Probabilistic modeling with Bayesian networks for automatic speech recognition. *Australian Journal of Intelligent Information Processing*, 253 - 260.

Appendices

A1: Preprocessing

```
claims <- read.xlsx("~/Downloads/Chronic_Claims_Details_Actisure_updated.xlsx", sheet =3)
df2 <- read.xlsx("~/Downloads/Chronic_Claims_Details_Actisure_updated.xlsx", sheet =2)
df3 <- read.xlsx("~/Downloads/Chronic_Claims_Details_Actisure_updated.xlsx", sheet =1)

#claims<-rbind(df,df2,df3)
Chronic_background <- read.xlsx("~/Downloads/Chronic_Background_Details.xlsx", sheet =1)

claims[1,]
```

Sample data

```{r warning=FALSE}
claims<-claims[,c(-1,-2,-3)]


claims[1:5,]
```

Total number of categories in the dataset:

```{r warning=FALSE, echo=FALSE}
cat("Total Claims \t",nrow(claims))
cat("\nUnique BENEFICIARYID \t", length(unique(claims$BENEFICIARYID)))
cat("\nUnique FIRSTDIAGNOSIS \t", length(unique(claims$FIRSTDIAGNOSIS)))
cat("\nUnique SECONDDIAGNOSIS \t", length(unique(claims$SECONDDIAGNOSIS)))
cat("\nUnique AFFECTEDSYSTEMDESC \t", length(unique(claims$AFFECTEDSYSTEMDESC)))
cat("\nUnique RISKGROUPDESC \t", length(unique(claims$RISKGROUPDESC)))
cat("\nUnique CLAIMYears \t", length(unique(claims$CLAIMDATE)))
cat("\nUnique TREATMENTDESCRIPTION \t", length(unique(claims$TREATMENTDESCRIPTION)))
cat("\nUnique SECTIONCODEDESC \t", length(unique(claims$SECTIONCODEDESC)))
cat("\nUnique PROVIDERNAME \t", length(unique(claims$PROVIDERNAME)))

```
```{r}
claims$FIRSTDIAGNOSIS<-as.factor(claims$FIRSTDIAGNOSIS)
count(claims, vars=c("FIRSTDIAGNOSIS","BENEFICIARYID"))
```

```

Unique CLAIMDATE
```{r warning=FALSE, echo=FALSE}
as.data.frame(table(claims$FIRSTDIAGNOSIS))
```
Unique FIRSTDIAGNOSIS
```{r warning=FALSE, echo=FALSE}
as.data.frame(table(claims$FIRSTDIAGNOSIS))
```

Unique SECONDDIAGNOSIS
```{r warning=FALSE, echo=FALSE}

new_DF <- claims[is.na(claims$FIRSTDIAGNOSIS),]
nrow(new_DF)
```

Unique AFFECTEDSYSTEMDESC
```{r warning=FALSE, echo=FALSE}
as.data.frame(table(claims$AFFECTEDSYSTEMDESC))[1:10,]
```

Unique RISKGROUPDESC
```{r warning=FALSE, echo=FALSE}
as.data.frame(table(claims$RISKGROUPDESC))[1:10,]
```

Unique TREATMENTDESCRIPTION
```{r warning=FALSE, echo=FALSE}
as.data.frame(table(claims$TREATMENTDESCRIPTION))[1:10,]
```

Unique SECTIONCODEDESC
```{r warning=FALSE, echo=FALSE}
as.data.frame(table(claims$SECTIONCODEDESC))
```

Unique PROVIDERNAME
```{r warning=FALSE, echo=FALSE}
as.data.frame(table(claims$PROVIDERNAME))[1:10,]
```



```{r warning=FALSE, echo=FALSE}

```
dff<-merge(x = claims, y = Chronic_background, by = "BENEFICIARYID", all.x = TRUE)
nrow(dff)

```
Total number of clients in the dataset
```{r warning=FALSE, echo=FALSE}
nrow(Chronic_background)
as.data.frame(table(Chronic_background$GENDER))

```
```{r warning=FALSE, echo=FALSE}
dff<-merge(x = claims, y = Chronic_background, by = "BENEFICIARYID", all.x = TRUE)
dff$GENDER <- as.character(dff$GENDER)
dff$GENDER[dff$GENDER == "F"] <- "FEMALE"
dff$GENDER[dff$GENDER == "Female"] <- "FEMALE"
dff$GENDER[dff$GENDER == "female"] <- "FEMALE"
dff$GENDER[dff$GENDER == "Ms"] <- "FEMALE"
dff$GENDER[dff$GENDER == "Male"] <- "MALE"
dff$GENDER[dff$GENDER == "male"] <- "MALE"
dff$GENDER[dff$GENDER == "M"] <- "MALE"

#Chronic_background[Chronic_background$GENDER == 'F','GENDER'] <- "FEMALE"
```
Though a majority of clients are born in the year 1980, they are distributed amoing the years with the youngest being 2016
```{r warning=FALSE, echo=FALSE}
year<-substring(Chronic_background$DOB,1,4) #

year_Df<-as.data.frame(table(year))
year_Df[order(-year_Df$Freq),][1:10,]
```
```
library(tidyverse)
freqclaim<-claims %>%
  select(BENEFICIARYID) %>%
  group_by(BENEFICIARYID) %>%
  summarise(claims = n()) %>%
  filter(claims > 5)

freqclaim

```

```{r}
freqclaim<-merge( freqclaim, Chronic_background,by="BENEFICIARYID")
freqclaim$DOB<- substring(freqclaim$DOB,1,4)

```

```

```{r}
freqclaim<-freqclaim[,c(-5,-6,-7,-9,-11)]


freqclaim[1:5,]


```

```{r}
freqclaim$DOB <- cut(x = as.numeric(freqclaim$DOB), breaks = c(1900, 1940, 1950, 1960, 1970, 1980,1990,2000,2010,2020))

freqclaim[1:5,]
```
dff<-merge(x = claims, y = Chronic_background, by = "BENEFICIARYID", all.x = TRUE)
dff$GENDER <- as.character(dff$GENDER)
dff$GENDER[dff$GENDER == "F"] <- "FEMALE"
dff$GENDER[dff$GENDER == "Female"] <- "FEMALE"
dff$GENDER[dff$GENDER == "female"] <- "FEMALE"
dff$GENDER[dff$GENDER == "Ms"] <- "FEMALE"
dff$GENDER[dff$GENDER == "Male"] <- "MALE"
dff$GENDER[dff$GENDER == "male"] <- "MALE"
dff$GENDER[dff$GENDER == "M"] <- "MALE"

wide<-bn.claims[,-1]
wide <- reshape(wide,
        idvar = "BENEFICIARYID",
        timevar = "AFFECTEDSYSTEMDESC",
        direction = "wide")

wide=as.matrix(wide)
wide[is.na(wide)] <-"NULL"
wide=as.data.frame(wide)

A2: Temporal Abstraction

library("bnstruct")
library(dplyr)
library(tidyr)
```

```
library(ggplot2)
library(tidyverse)
library(openxlsx)
library(ggplot2)
library(tidyverse)

claims.data<-read.csv("~/Dropbox/claims.csv")
claims.d<-unique(claims.data)

newdf <- claims.d[,c(-6,-8)] %>%
  group_by(BENEFICIARYID) %>%
  arrange(CLAIMDATE) %>%
  mutate(replicate=seq(n())) %>%
  mutate(freq=max(replicate)) %>%
  filter (freq > 3)

time1 <- newdf %>%
  filter (freq > 3) %>%
  filter (replicate == 1)

write.csv(time1,"~/Dropbox/time1.csv")

time2 <- newdf %>%
  filter (freq > 3) %>%
  filter (replicate == 2)

write.csv(time2,"~/Dropbox/time2.csv")

time3 <- newdf %>%
  filter (freq > 3) %>%
  filter (replicate == 3)

write.csv(time3,"~/Dropbox/time3.csv")


time4 <- newdf %>%
  filter (freq > 3) %>%
  filter (replicate == 4)

write.csv(time4,"~/Dropbox/time4.csv")

t12<-merge(time1, time2,  by="BENEFICIARYID")
t123<-merge(t12, time3,  by="BENEFICIARYID")
t1234<-merge(t123, time4,  by="BENEFICIARYID")
write.csv(t1234,"~/Dropbox/t1234.csv")
```

A3: Model Building

```
# Create a Bnstruct data set
```

```
datas <- BNDataset(data = cdatamat, discreteness = c(T,T,T,T,
                                        T,T,T,T,
                                        T,T,T,T,
                                        T,T,T,T),   num.time.steps = 4, na.string.symbol=NA,
variables =
                 c("FIRSTDIAGNOSIS","RISKGROUP","AGE","GENDER",
                   "FIRSTDIAGNOSIS_1","RISKGROUP_1","AGE_1","GENDER_1",
                   "FIRSTDIAGNOSIS_2","RISKGROUP_2","AGE_2","GENDER_2",
                  "FIRSTDIAGNOSIS_3","RISKGROUP_3","AGE_3","GENDER_3"),
               node.size =c(228,33,6,2,224,36,6,2,227,35,6,2,243,36,6,2) )

# Define layers
layerstruct=matrix(0, 3, 3)

layerstruct[1,2]<-1
layerstruct[2,3]<-1
layerstruct[2,2]<-1
layerstruct[3,3]<-1
layerstruct[3,2]<-1
# Define Edges
lay<-matrix(0,16,16)

lay[1,2]<-1
lay[4,1]<-1
lay[3,1]<-1
lay[7,5]<-1
lay[8,5]<-1
lay[5,6]<-1
lay[11,9]<-1
lay[12,9]<-1
lay[9,10]<-1
lay[1,5]<-1
lay[5,9]<-1
lay[13,14]<-1
lay[15,13]<-1
lay[16,13]<-1
lay[9,13]<-1

# Build model
layers <- c(3,2,1,1,3,2,1,1,3,2,1,1,3,2,1,1)

dbnn <- learn.dynamic.network(datas, num.time.steps = 4,algo = "mmhc", scoring.func
= "BIC", layering=layers,layer.struct=layerstruct, mandatory.edges=lay,custom=lay )
```

A4: Using Inference Engine

Example 1

```{r}
```

```
#"FIRSTDIAGNOSIS","RISKGROUP","AGE","GENDER",
obs <- list("observed.vars" = c(1,3,4,5,7,8,11,12),
 "observed.vals" = c(52,5,2,62,5,2,5,2))
engine <- InferenceEngine(dbnn)
engine <- belief.propagation(engine, obs)
new.net <- updated.bn(engine)```

```{r}
ccc<-cpts(new.net)
dignosis1<-ccc[[13]]
str(dignosis1)

ord<-order(matrix(dignosis1[56,,5,2]),decreasing=TRUE)[1:4]
ord
dignosis1[56,ord,5,2]
```
Example 2
```{r}
#"FIRSTDIAGNOSIS","RISKGROUP","AGE","GENDER",
obs <- list("observed.vars" = c(1,3,4,5,7,8,11,12),
 "observed.vals" = c(87,5,2,92,5,2,5,2))
engine <- InferenceEngine(dbnn)
engine <- belief.propagation(engine, obs)
new.net <- updated.bn(engine)

```

```{r}
ccc<-cpts(new.net)
dignosis1<-ccc[[13]]
str(dignosis1)

ord<-order(matrix(dignosis1[84,,5,2]),decreasing=TRUE)[1:4]
ord
dignosis1[84,ord,5,2]
```

Example 3
```{r}
#"FIRSTDIAGNOSIS","RISKGROUP","AGE","GENDER",
obs <- list("observed.vars" = c(1,3,4,5,7,8,11,12),
 "observed.vals" = c(87,4,1,62,4,1,4,1))
engine <- InferenceEngine(dbnn)
engine <- belief.propagation(engine, obs)
new.net <- updated.bn(engine)

```
```

```{r}
ccc<-cpts(new.net)
dignosis1<-ccc[[13]]
str(dignosis1)

ord<-order(matrix(dignosis1[80,,4,1]),decreasing=TRUE)[1:4]
ord
dignosis1[80,ord,4,1]
```

Example 4

```{r}
#"FIRSTDIAGNOSIS","RISKGROUP","AGE","GENDER",
obs <- list("observed.vars" = c(1,3,4,5,7,8,11,12),
 "observed.vals" = c(102,4,1,112,4,1,4,1))
engine <- InferenceEngine(dbnn)
engine <- belief.propagation(engine, obs)
new.net <- updated.bn(engine)

```

```{r}
ccc<-cpts(new.net)
dignosis1<-ccc[[13]]
str(dignosis1)

ord<-order(matrix(dignosis1[107,,4,1]),decreasing=TRUE)[1:4]
ord
dignosis1[107,ord,4,1]
```

Example 5

```{r}
#"FIRSTDIAGNOSIS","RISKGROUP","AGE","GENDER",
obs <- list("observed.vars" = c(1,3,4,5,7,8,11,12),
 "observed.vals" = c(52,3,1,196,3,1,3,1))
engine <- InferenceEngine(dbnn)
engine <- belief.propagation(engine, obs)
new.net <- updated.bn(engine)

```

```{r}
ccc<-cpts(new.net)
dignosis1<-ccc[[9]]
str(dignosis1)
```

```
ord<-order(matrix(dignosis1[196,,3,1]),decreasing=TRUE)[1:4]
ord
dignosis1[196,ord,3,1]
```