Predicting the Risk of Chronic Disease:
A Framework Based on Graph Theory and Social Network Analysis

Mohammad Ariful Alam Khan

A thesis submitted in fulfilment of the requirements for the degree of
Doctor of Philosophy

Faculty of Engineering and IT
The University of Sydney

9 March 2017
Abstract

Chronic diseases and conditions are the leading causes of death worldwide. In 2005, approximately 35 million people died from chronic diseases, making them responsible for 60% of all mortalities. Despite the popular belief that chronic diseases affect mostly high-income countries and older people, the reality gives quite a different picture: 80% of deaths from chronic diseases occur in low- and middle-income countries, and these are mostly concentrated among the poor. Almost half of the deaths occur at an age under 70 years. The premature mortality and reduced quality of life for chronic patients not only results in an epidemiological burden, but also exerts a significant impact on national economies. Therefore, chronic disease prevention and management has been a major concern for governments and related international organisations.

This thesis narrows on a particular scope within the greater paradigm of chronic disease prevention and management, which is to understand chronic disease progression and then predict risk based on this understanding. The motivation for this project comes from the fact that a significant proportion of potentially preventable hospital admissions are due to chronic disease. In many cases, patients are unaware of their chronic conditions, and only when they are admitted to hospital are these chronic conditions discovered as a secondary diagnosis. This results in further complications, longer length of stay and an increased burden on limited healthcare resources. Had they been diagnosed as chronic patients before, all of these complications could have been averted. However, traditional methods of clinical diagnosis and regular monitoring of a large population are often resource-intensive in terms of available clinical provisions and economic capability. One potential alternative comes from a data mining perspective on healthcare information systems, more specifically hospital admission data, which carries rich semantic information about the patients’ overall health status and diagnosis information in the form of standardised codes. This vast amount of systematically generated data can help us to understand the disease footprints left by chronic patients. The understanding can then
be utilised to evaluate the health status of another population. At present, very limited work has been done in realising these particular potentials of this administrative data.

This thesis, therefore, asks the questions: how administrative data can generate knowledge that can help us to understand chronic disease pathways? Is it possible to reasonably predict chronic disease risk by leveraging our understanding of disease progression? This thesis attempts to answer these questions by presenting a framework. The framework has two major parts: (i) to understand and represent the progression of a particular chronic disease, and (ii) to develop a model based on that understanding to predict chronic disease risk for non-chronic patients. We utilised graph theory and social network analysis for both of these parts. For the first part, we propose the concept of a ‘baseline network’ that can effectively model chronic disease comorbidities and their transition patterns, thereby representing the chronic disease progression. We further take the attribution effect of the comorbidities into account while generating the baseline network; that is, we not only look at the pattern of disease in the chronic disease patients, but also compare them with that of non-chronic patients to understand which comorbidities are more responsible for leading to the chronic disease pathway. For the second part, we used this ‘baseline network’ to compare against the individual health trajectories of non-chronic patients. For matching the networks, we proposed several graph theory and social network-based methods. These methods look at multiple parameters, including the prevalence of the comorbidities, transition patterns and frequencies, clustering membership, and demographic and behavioural factors such as age, sex and smoking. Individual risk scores against each of these parameters are then merged to generate the single prediction score that purports to represent the risk of future chronic disease for current non-chronic patients.

We implemented the framework on administrative data drawn from the Australian healthcare context and chose type 2 diabetes as the chronic disease for the model to predict. The overall dataset contained approximately 1.4 million admission records from .75 million patients, from which we filtered and sampled the records of 2,300 diabetic and
2,300 non-diabetic patients. We followed an exploratory approach during the implementation to understand the relative contributions of different types of parameters in the prediction. We utilised three different predicting modelling methods—regression, parameter optimisation and tree classification—all of which gave the highest ranking to the graph theory-based ‘comorbidity prevalence’ and ‘transition pattern match’ scores. This proves the effectiveness of our proposed network theory-based measures. We also explored the effect of using limited and selected comorbidities compared to using fixed-length International Classification of Diseases (ICD) codes in representing the actors or nodes of the baseline network. Fixed-length ICD codes, together with tree classification, showed the highest accuracy. Overall, the framework showed prediction accuracy within 90% to 95% in identifying the future risk of type-2 diabetes. This is a significant improvement over the other similar prediction methods. Further, we critically analysed the effect of using different predictive models, attribution adjustment and ICD code collapse on performance. This provides direction on which specific method settings to use based on the specific chronic disease and type of healthcare dataset, thereby making the overall framework more generic and versatile.

The framework proposed in the thesis can potentially be useful for stakeholders including governments and health insurers to identify cohorts of patients at high risk of developing chronic disease. Adopting a rigorous early intervention and prevention policy targeted at those patients can potentially divert them from the chronic disease pathway, and reduce healthcare costs from both provider and consumer perspective.
Declaration of Originality

I hereby declare that this submission is my own work and that, to the best of my knowledge and belief, it contains no material previously published or written by another person nor material which to a substantial extent has been accepted for the award of any other degree or diploma of the University or other institute of higher learning, except where due acknowledgement has been made in the text.

Signed: Mohammad Ariful Alam Khan On: 09 / 03 / 2017
Preface

We started this research with the focus to quantify the longitudinal nature of complex and dynamic networks. To do that, we developed a set of measures during the first year. We combined them into an analytical framework and started to implement them on longitudinal datasets like obesity research data, collaboration network etc. At the same time, we were looking for a more robust context to extend the framework that can have a practical implication. By the end of 2014, we got the opportunity to collaborate with Capital Markets Cooperative Research Centre (CMCRC) who became our industry partner. CMCRC provided us the access to a large private health data after all the parties sorted out rigorous security, privacy and ethical requirements. Access to this health data opened up a vast research scope in front of us. This was the turning point that shaped the overall research questions and the way this thesis is presented.

The target audience of this research are likely to be data and network scientists, healthcare related policy makers and analytics as the research lies between the broad disciplines of data mining, health informatics, complex systems and network science. Keeping this in mind, we tried to present it comprehensibly and avoided technical terms as much as possible. Also, before ‘introduction’ we added the most frequent terminologies used in the thesis. A comprehensive list of terms, abbreviations and related background information are included later in appendix. The total length of the thesis is approximately 61,000 words excluding appendices and references.

This research used de-identified personal health data of patients containing information like age, sex, hospital admission summary etc. Required human ethics approval was obtained from the Human Research Ethics Committee of University of Sydney under the project number 2015/824. A copy of the approval letter is attached in the Appendix D.

This thesis was first submitted on 31 August, 2016. Revised version was submitted for final lodgement on 9 March 2017.
Acknowledgements

After years of relentless work, my PhD candidature is finally coming to an end with this final lodgement of this thesis. Looking in retrospect, since 2013 until now, I should say, it had been a wonderful journey. Of course, it was not always a pleasant. There were times when I spent hours staring at my own code, trying to find that tiny software bug messing up the whole result. There were times when things seemed impossible to solve. Yet, at the end, it all worked out and the hours I spent looking for those bugs – they were all worth the result.

Without the inspiration and sincere help of several people, this thesis would not be possible. Among them, I should first express my sincere gratitude towards my supervisor Dr Shahadat Uddin. The amount of guidance, support, motivation and cooperation that he has selflessly extended, from the day one till the very last day, is truly appreciable. Second, I want to thank my industry supervisor Dr Uma Srinivasan for her mentorship and evergreen enthusiasm in explaining the complex subjects of Australian health system, policies and administrative data collection process. I would also like to thank Prof. Liaquat Hossain, for his support during the first year of my candidature.

I would like to thank all the related staffs at Capital Markets Cooperative Research Centre (CMCRC) for their support. My heartfelt gratitude to David and Dr Federico for looking over the Health Market Quality (HMQ) Program and the constant support on various issues. Thank you, Allan, Jawad and Alastair for your timely support in data transfer and IT issues. Thank you, Amir, for your help with the R codes; Eugenio, Panha and Eddy for making the workplace funnier. Thanks to other HMQ students – Kelsey, Jo and Maneesh for your help and insights. From the University of Sydney, I want to express my gratitude to Nazim for his support in conducting the research. I would also like to thank Sagar for helping me out with frequent help and advice on different academic issues. I am also grateful to my housemates for their cooperation, support that had my life in Sydney much easier.
I wish to acknowledge the financial support and the data analytics platform provided by the Capital Markets Cooperative Research Centre under the Health Market Quality Program and our industry partner HAMB Systems Ltd.

I want to thank all the office staffs at the University of Sydney - Maria, Daniela, Lorraine, Dami, John barber and Gary for their tremendous support.

This thesis was edited by Elite Editing, and editorial intervention was restricted to Standards D and E of the Australian Standards for Editing Practice. I would like to thank them for rendering the service efficiently.

At the end, I want to thank my family – my mother, father and brother for their constant support. Living far away from home, this would not be possible without your inspiration. The teaching and guidance that you gave me since childhood, has brought me up to this point. Words are probably not enough to express my gratitude. So, I will just say, thank you for everything you have done for me.

Update on 9 March 2017: Two months ago, I married to a beautiful girl. She reminds me that life is wonderful even with the task of doing a PhD. So, thank you, Era – for all the love and support during these last few weeks of revision.
Dedication

To my wonderful family:
Amma, Abba, Bhaiya and dear Era
I love you all

And to all the scientists around the world
working hard for making lives better
this is my tiny contribution for your endeavour
List of Publications


5. Uddin, S., **Khan, A.** and Baur, L.A. A framework to explore the knowledge structure of multidisciplinary research fields. *PloS one*, 10 (4). (Impact Factor: 3.54)


7. Uddin, S., **Khan, A.** and Piraveenan, M. Administrative claim data to learn about effective healthcare collaboration and coordination through social network. in *Hawaii International Conference on System Sciences (HICSS-48)*, (2015), IEEE, 3105-3114.

9. Uddin, S., Piraveenan, M., Khan, A. and Amiri, B. Conceptual quantification of the
dynamicity of longitudinal social networks. in Social Computing (SocialCom), 2013

Book chapters

1. Srinivasan, U., Khan, A., and Uddin, S. Network Analytics to Enable Decisions in
Table of Contents

Abstract ................................................................................................................................. ii
Declaration of Originality ................................................................................................... v
Preface .................................................................................................................................. vi
Acknowledgements ............................................................................................................. vii
List of Publications ............................................................................................................... x
Table of Contents ............................................................................................................... xii
List of Tables ..................................................................................................................... xvi
List of Figures ................................................................................................................... xvii
List of frequently used Terms and Abbreviations ............................................................. xviii

Chapter 1: Introduction ...................................................................................................... 1
  1.1 Objectives ..................................................................................................................... 1
  1.2 Background and Motivation ......................................................................................... 3
    1.2.1 Burden of chronic diseases ................................................................................... 3
    1.2.2 Modern health system and availability of healthcare data .................................... 6
    1.2.3 Policy-level changes in the healthcare industry .................................................... 9
  1.3 Key Research Questions ............................................................................................ 11
  1.4 Approach .................................................................................................................... 12
  1.5 Contribution of the Study ........................................................................................... 13
    1.5.1 Theoretical contribution ...................................................................................... 13
    1.5.2 Methodological contribution ............................................................................... 14
    1.5.3 Knowledge translation for the health policy maker ............................................ 15
  1.6 Thesis Structure .......................................................................................................... 16

Chapter 2: Background and Related Research ............................................................... 19
  2.1 Evolution of Healthcare Informatics .......................................................................... 21
    2.1.1 Prognosis: from ancient times to Greek medicine .............................................. 21
    2.1.2 Romans, Arabs and medieval Europe .................................................................. 23
    2.1.3 Renaissance and the industrial revolution ........................................................... 25
    2.1.4 Introduction of computer and medical informatics ............................................. 26
    2.1.5 Electronic health records and administrative data .............................................. 28
  2.2 Statistical Methods of Disease Risk Prediction ......................................................... 29
  2.3 Data Mining-Based Disease Prediction ..................................................................... 31
    2.3.1 Score-based models ............................................................................................. 32
    2.3.2 Collaborative filtering ......................................................................................... 35
  2.4 Networked Approach in Healthcare ........................................................................... 38
  2.5 Social Network Analysis in Medical Informatics ...................................................... 41
  2.6 Limitations of Existing Theories ............................................................................... 44
Chapter 5: Results and Discussions

5.1 Descriptive Statistics
5.2 Criteria for Cohort Selection
  5.2.1 Criteria for selecting the diabetic cohort
  5.2.2 Criteria for selecting the non-diabetic cohort
5.3 Characteristics of the Diabetic and Non-diabetic Cohorts
5.4 Code Range Selection
  5.4.1 Selective comorbidities based on the Elixhauser index
  5.4.2 ICD code collapse
  5.4.3 Selection for socio-demographic factors
5.5 Comparison Across Three Baseline Networks
5.6 Findings for Attribution Effects
  5.6.1 Prevalent comorbidities
  5.6.2 Prevalent transition patterns
  5.6.3 Clustering membership in baseline network
5.7 Parameter Estimation
  5.7.1 Influence of risk factors from parameter estimation model
  5.7.2 Influence of risk factors from logistic regression
  5.7.3 Influence of risk factors from binary tree classification method
5.8 Risk Assessment Results for Test Group
5.9 Validation and Comparison of Results
  5.9.1 Precision and recall result
  5.9.2 Effect of considering collapsed ICD codes
  5.9.3 Attribution effect
List of Tables

Table 2.1: Sample parameters to score against in score-based system ................................ 33
Table 2.2: Example of different scoring systems .................................................................. 35
Table 2.3: Comparison between different disease prediction models ................................. 46
Table 4.1: List of risk factors considered in the framework .................................................. 120
Table 4.2: Different types of outcomes from tests ................................................................. 126
Table 5.1: Overall statistics of the dataset ........................................................................... 129
Table 5.2: Basic statistics for research dataset ........................................................................ 133
Table 5.3: Prevalence of different types of diabetes ............................................................... 135
Table 5.4: List of ICD codes to determine frequent admissions ............................................ 139
Table 5.5: List of Elixhauser comorbidities used in the framework ........................................ 143
Table 5.6: Comorbidities and health conditions added to Elixhauser index ......................... 144
Table 5.7: Individual risk factors related to demographics and behaviour ............................ 148
Table 5.8: Network properties of three Baseline Networks (BN) ........................................... 150
Table 5.9: Top 10 most prevalent diseases for diabetic and non-diabetic patients ................. 155
Table 5.10: Top 10 most frequent progressions between comorbidities or health conditions in subsequent admissions according to the final baseline network .................... 159
Table 5.11: Cluster membership among comorbidities in the baseline network listed according to the prevalence rate ................................................................. 162
Table 5.12: Optimal weighting factors from parameter estimation model ............................ 169
Table 5.13: Coefficient values against risk factors in binary logistic regression model ......... 171
Table 5.14: Prediction results for different models for fold-0 ................................................ 177
Table 5.15: Comparison between different predictive models using collapsed ICD codes and selective comorbidities ................................................................. 184
Table 5.16: Performance comparison between inclusion and exclusion of attribute adjustment ................................................................................................................ 187
Table 5.17: Summary properties among different approaches used to implement the framework .................................................................................................................... 192
List of Figures

Figure 1.1: Comparison of percentage of total death from chronic and selected non-chronic causes in different regions (Cohn, 2007, Mathers, et al., 2003) .............. 4
Figure 1.2: Organisation ................................................................. 8
Figure 2.1: Organisation of literature review concepts ........................................ 20
Figure 2.2: Evolution of healthcare informatics and related fields with computer technology .................................................................................. 27
Figure 2.3: Sample network-based approach in disease or associated biomarker prediction ................................................................. 41
Figure 2.4: Different types of communication pattern .............................................. 50
Figure 2.5: Different funding sources for total health expenditure during 2013–14 60
Figure 2.6: Organisation of Australian health system ................................................ .. 61
Figure 3.1: Overview of data de-identification process .............................................. 68
Figure 3.2: Simplified version of Entity-Relationship (E-R) diagram ....................... 70
Figure 3.3: Additional entities derived from the database ........................................... 75
Figure 3.4: Health trajectory of an individual chronic disease patient ...................... 81
Figure 3.5: Functional components of the framework .............................................. 85
Figure 3.6: Summary of the framework .................................................................. 90
Figure 3.7: Individual disease network and attributes .............................................. 94
Figure 3.8: Degree based centrality measures in a baseline network ....................... 98
Figure 3.9: Cluster detection in a sample baseline network ..................................... 100
Figure 3.10: Process of cohort selection ............................................................... 107
Figure 4.1: Construction of intermediate time-detailed network ............................ 109
Figure 4.2: Merging of time gap information to convert a time-detailed graph into a baseline network ................................................................. 111
Figure 4.3: Example process of attribution adjustment for nodes ............................ 114
Figure 4.4: Process of choosing optimum threshold value and accuracy .................. 121
Figure 5.1: Monthly admission statistics for entire dataset .................................... 131
Figure 5.2: Comorbidity pattern within different forms of diabetes ....................... 135
Figure 5.3: Age distribution of the diabetic and non-diabetic cohorts .................... 140
Figure 5.4: Time between first and last considered admission for diabetic and non-diabetic cohorts ............................................................................. 141
Figure 5.5: The three baseline networks and their network and node level comparison ................................................................. 149
Figure 5.6: Top 10 comorbidities and conditions that attributed most for diabetes patients .................................................................................. 157
Figure 5.7: Baseline network after attribute adjustment ......................................... 160
Figure 5.8: Parameter estimation component of the framework ................................ 165
Figure 5.9: Scatterplot matrices for the risk factors .............................................. 166
Figure 5.10: Classification tree generated from the binary tree classification method (output from fold-0 iteration) ................................................................. 172
Figure 5.11: Parameter importance over 10 folds in tree classification method (behaviour risk factor not shown) ........................................................................ 174
Figure 5.12: Overall risk scores of the test cohorts, predicted outcomes and corresponding grouping according to the parameter estimation model .......... 180
Figure 5.13: Prediction accuracy (%) for all folds for all three methods ...................... 182
Figure 5.14: Precision and recall measures for all folds for all three methods ................ 183
Figure 6.1: Different groups of members having different health status for a typical health fund and their shifts over time ............................................. 198
<table>
<thead>
<tr>
<th>Term</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>AIHW</td>
<td>Australian Institute of Health and Welfare</td>
</tr>
<tr>
<td>Baseline Network</td>
<td>A network representation of the health trajectory of chronic disease or any particular group of patients.</td>
</tr>
<tr>
<td>BMI</td>
<td>Body Mass Index</td>
</tr>
<tr>
<td>CDMP</td>
<td>Chronic Disease Management Program</td>
</tr>
<tr>
<td>Chronic Disease</td>
<td>A broad set of illnesses and health conditions that are non-communicable and persist over a long period of time, such as – diabetes, cancer, cardiovascular diseases etc.</td>
</tr>
<tr>
<td>Comorbidity</td>
<td>Diseases that occur together. In clinical coding context, the disease chiefly responsible for the admission is called principal diagnosis. Other existing diseases or conditions are recorded as secondary diagnoses and are called comorbidities.</td>
</tr>
<tr>
<td>COPD</td>
<td>Chronic obstructive pulmonary disease</td>
</tr>
<tr>
<td>Disease Progression</td>
<td>The occurrence of different diseases of comorbidities over time and the transitions between them.</td>
</tr>
<tr>
<td>DRG</td>
<td>Diagnosis-Related Group</td>
</tr>
<tr>
<td>E-R</td>
<td>Entity-Relationship</td>
</tr>
<tr>
<td>GP</td>
<td>General practitioner</td>
</tr>
<tr>
<td>HCP</td>
<td>Hospital Casemix Protocol</td>
</tr>
</tbody>
</table>

1 A more extensive Glossary and Acronyms are given in the Appendix A and B.
Health Fund  (In thesis context) Organisations providing private health insurance.

Health Trajectory  Evolution of the health status of a single or group of patients over time. In this research context, we used diseases (i.e., ICD codes) as the indicator of health status or health trajectory.

ICD  International Classification of Diseases

ICU  Intensive care unit

MBS  Medicare Benefits Schedule

NDSS  National Diabetes Services Scheme

NSW  New South Wales

PBS  Pharmaceutical Benefits Scheme

SNA  Social network analysis

US  United States

WHO  World Health Organization
Chapter 1: Introduction

In this chapter, we present a broad-level overview of the whole thesis. We begin by stating the primary objective of our research. We then briefly discuss the contextual background and the primary motivations that led the research project. Following the background and motivation, we list the research questions and approach followed in the thesis. Finally, we discuss the major contributions of the research, followed by an overview of the thesis that is presented in the subsequent chapters. As our research is based on the Australian healthcare system, we try to provide statistics focused on Australia and countries dependent on similar policies to make the thesis more comprehensible and provide relevant context.

1.1 Objectives

This thesis has two broad goals: (1) to understand and represent the progression of chronic disease, and (2) to develop a model based on this understanding to predict chronic disease risk for non-chronic patients. To achieve them in real-world healthcare data as context, this thesis develops a set of methodologies which are mostly based on graph theory and social network analysis. Together, these methodologies provide a generic and systematic model to understand chronic disease progression and assess future risks. Throughout this thesis, the term ‘framework’ is used to collectively refer to this set of methodologies.

Chronic diseases are a broad set of different illnesses and health conditions that are non-communicable and persist over a long period of time. The four most prevalent chronic disease groups are: cardiovascular diseases (e.g., stroke, heart attacks), cancers, chronic obstructive pulmonary disease (COPD) and diabetes. These diseases generally do not have preventive vaccines or definitive cures once the patient has developed the condition. Previous research has shown that chronic diseases often occur together as comorbidities, because they share similar underlying causes and risk factors. As a result, the progression
towards a particular chronic disease often comes from a complex interplay and transition between different diseases or comorbidities. We define these interactions and transitions between comorbidities leading to a particular chronic disease as ‘disease progression’. The first part of our framework aims at understanding this disease progression—that is, the comorbidities and their prevalence and transition patterns over time. We propose a networked representation called a ‘baseline network’ for understanding this disease progression.

We use the administrative claim dataset that is routinely generated when patients are admitted to hospitals. This dataset contains all disease-related information for admitted patients along with timestamps, thus enabling us to obtain a snapshot of their health trajectory. This information is used in the first part of the framework to generate a baseline network to understand the progression of the chronic disease. For this thesis, the chronic disease we focus on is type 2 diabetes. This is a progressive condition in which the body becomes resistant to the normal effects of insulin (a hormone) and/or gradually loses the capacity to produce enough insulin in the body’s pancreas (Thomas, 2015).

In the second part of the framework, we make a prediction model by utilising the empirical understanding gained from the previous part, i.e., from the baseline network of the comorbidities that lead to type 2 diabetes. The goal is to propose a set of graph theory and social network-based measures to compare the baseline network with the health trajectory of a non-diabetic (i.e., non-chronic) patient. These measures are then used to predict the overall risk of diabetes for that patient.

In addition to the primary goals, the framework adopts an explorative approach to developing secondary understanding throughout the research. This is necessary because hospital administrative datasets have rarely been used in prior studies that properly align with our research goals. Therefore, we want to explore different alternative methods within the framework to test their effectiveness. For example, we employ three different prediction models in the second part of the thesis to compare their performance and discuss individual benefits and pitfalls. Further, we consider several parameters based on
demographic and behavioural risk factors. Parameters based on the baseline network and
disease cluster matching are also considered. The goal is to understand and quantify these
parameters’ influence in diabetes prediction and discuss why some parameters are better
predictors than others. It is hoped that the explorative comparison between different
approaches within the framework will make it more flexible in its applicability to other
chronic diseases and dataset contexts.

1.2 Background and Motivation

1.2.1 Burden of chronic diseases

Since antiquity, major causes of human death have been mostly concerned with childhood
mortality, infectious disease, malnutrition, epidemics and physical injury. Modern
scientific and medical advancement has drastically reduced deaths due to these causes
through vaccines, antibiotics, and improved medication and diagnostic techniques.
Conversely, chronic non-communicable diseases have become the leading causes of death
in most parts of the world. In 2012, 38 million people died of chronic diseases worldwide,
representing 68% of all death (World Health Organization, 2012). Similar statistics are also
observed across countries within last decade. For instance, in United States (US) as of
2012, 117 million people (about half of all adults) had one or more chronic conditions
(Ward, et al., 2014), 90% of all deaths in Australia in 2011 were due to chronic disease
(AIHW, 2014). While it is often assumed that chronic diseases are only prevalent problems
for richer countries, statistics paint a completely different picture: 80% of deaths from
chronic diseases occur in low- and middle-income countries, mostly concentrated among
the poor. One of the reasons for this trend is that many low- and middle-income countries
are undergoing economic and nutritional transition (Popkin, et al., 2012), resulting in a
higher prevalence of obesity, which is a common risk factor for many chronic diseases. As
a result, many developing countries have to deal with infectious diseases and malnutrition
side by side with the alarmingly increasing prevalence of obesity and chronic diseases.
Figure 1.1 shows the evidence that deaths related to chronic disease are leading in most regions of the world compared to deaths from non-chronic diseases, according to research from the WHO (Cohn, 2007, Mathers, et al., 2003).

Another common misconception is that chronic diseases occur mostly in the elderly. In reality, almost half of all deaths from chronic diseases occur at an age under 70 years, when people are still in their productive years. Premature mortality is, however, not the only concern. As has been mentioned above, across the world countries are undergoing economic transition. Rising incomes, falling food prices, improved supply chains and urbanisation have led to global changes in diet, overweight, obesity and physical inactivity across all age groups (Alwan, 2011). As a result, early onset of chronic conditions is ever on the rise (Kopelman, 2000). The potential loss of productivity, illness, disability, death and healthcare costs thus put a significant strain on the economy.
The primary motivation for this thesis has come from the efforts to prevent chronic diseases exerted by government, stakeholders, healthcare providers and patients. The good news is that many chronic diseases are preventable. For example, 80% of premature heart disease, stroke and type 2 diabetes and 40% of cancer cases are preventable (WHO, 2005). Healthy and careful lifestyle choices can prevent the onset of complications or manage them once they are developed. However, the problem is that many chronic diseases, especially type 2 diabetes, develop slowly without explicit symptoms; thus the patients may remain unaware of their condition until it is too late (Gregg, et al., 2004, Rathmann, et al., 2003). Moreover, when patients are admitted to hospitals for treatments or surgical procedures, they are often diagnosed with chronic disease such as type 2 diabetes (Taubert, et al., 2003) as comorbidity, not the primary reason for admission. There is substantial evidence to show that patients with diabetes take longer to recover, have higher mortality and contract more secondary infections, and in general there is a substantial increase in the cost of care to achieve the desired health outcomes (Harris, 1993, Kapur, et al., 1999, Lauruschkat, et al., 2005, MacKenzie, et al., 1989, Tenenbaum, et al., 2000, Umpierrez, et al., 2002).

The growing evidence, therefore, suggests the need for more accurate diagnosis of chronic diseases. The intuitive way is to achieve this need is to implement a healthcare policy that encourages rigorous and timely screening for chronic diseases. However, mass screening is often expensive and difficult to implement, given the scarcity of clinical resources and related infrastructure in many countries. Therefore, a cost-effective option could be to target only a portion of the population that are at high risk based on their previous medical conditions, as many chronic diseases, as we have observed, can demonstrate a comorbidity pattern leading up to the disease. There are two challenges: first, we need to understand the comorbidity pattern, and, based on that pattern, identify the cohort most at risk. Second, if we can achieve this, the at-risk cohort can be further screened or advised to take preventive measures, which can be a sustainable and cost-effective solution from both provider and consumer perspectives.
1.2.2 Modern health system and availability of healthcare data

Most countries today have well-defined health systems (often referred to as healthcare systems) that focus on providing affordable and quality healthcare services for their citizens. The WHO defined a health system as follows:

A health system consists of all organizations, people and actions whose primary intent is to promote, restore or maintain health. ... It includes, for example, a mother caring for a sick child at home; private providers; behaviour change programmes; vector-control campaigns; health insurance organisations; occupational health and safety legislation. It includes inter-sectoral action by health staff, for example, encouraging the ministry of education to promote female education, a well-known determinant of better health. (World Health Organization., 2007)

The broad goals of a health system are to improve health and health equity in ways that are responsive, financially fair, and make the best, or most efficient, use of available resources (World Health Organization., 2007, 2000).

Since the modernisation of computers and networking technology, the health systems of most countries have evolved and revolutionised the way healthcare services are delivered. As a consequence, modern health systems generate large amounts of data as the health consumer traverses the system, accessing services. The origin, type and purpose of healthcare-related datasets vary widely. Nevertheless, we can broadly categorise healthcare datasets into four groups based on the ways in which they are collected. The first type of health dataset is ‘survey data’ generated through systematic surveys, mostly conducted by researchers or other stakeholders, to understand trends in populations through specific tailored questions. Specifically, chronic diseases incur long-term burdens on healthcare systems as well as in the individual’s quality of life; therefore, many countries maintain an active registry of databases through longitudinal surveys of sections of the population. For an example, the organisation Diabetes Australia keeps an active registry of diabetes patients under the National Diabetes Services Scheme (NDSS) to monitor the prevalence of diabetes across Australia (Diabetes Australia, 2016). Another
example is the ‘45 and Up Study’, which regularly surveys a large population of Australians aged over 45 (SAX Institute, 2016).

The second type of health dataset, which is emerging as an important source of data, is ‘device-generated data’. These data are generated by electronic devices in an automated manner. For example, modern medical devices (e.g., ICU devices) may generate periodic vital statistics of patients, and these sources of data are often mined. Also, personal electronic devices (e.g., smart watches, wristbands, smartphone apps) may generate and share real-time health statistics.

The third type of health dataset is ‘medical records’. This type of dataset contains information related to a patient’s medical condition, including diagnosis and treatment. A medical record is created for each patient in most primary and tertiary healthcare settings and is mainly used by the providers. Typically, patients are assigned a unique identifier (e.g., Medical Record Number [MRN]) to link records created during subsequent consultations. As information technology has substantially improved globally, more countries have begun to implement the concepts of electronic healthcare, commonly known as e-health. As a result, individual medical records are shared, unified and often stored centrally. This has become a valuable resource for managing continuity of care, so that the various providers and emergency departments can have unified, up-to-date and consistent access to patients’ health records. The patient, in turn, can also have access to their complete medical history.

The last of the four datasets is the ‘administrative dataset’, which is collected for billing, quality assurance, auditing and other administrative purposes. Unlike the ‘medical records’ set, which is mostly maintained by healthcare providers to understand individual health conditions, administrative datasets contain more service-level information, such as fees and charges and summaries of provided services. These datasets are often shared across policy makers, funders, government agencies and other stakeholders, and can serve as rich sources of information. Figure 1.2 shows a snapshot of the four different types of healthcare datasets and their potential use by different stakeholders.
This research utilises the fourth type of dataset, i.e., ‘administrative dataset. Therefore, in this chapter and in the subsequent ones, we make particular effort to discuss the potential for knowledge discovery from the administrative dataset.

In Australia, as in several other countries, an important portion of administrative data consists of hospital admission and discharge details. Several applications that are illustrated in the following sections are based on data stored in such administrative records, which are kept in both public and private health sectors in Australia. The administrative record of a hospital-admitted patient contains: patient demographics; patient’s medical condition, represented as principal and secondary diagnosis ICD codes (Rizzo, et al., 2015); procedures performed, represented as MBS (AIHW, 2014) and International Classification of Diseases (ICD) procedure codes; and length of service.
provided (including accommodation and prosthetics). The billing information is of particular interest for private and public funders such as Medicare (Department of Human Services, 2016). As the records also contain information about patient demographics, time and duration of stay, diagnoses and treatment, they can be used to understand disease progression at a population level. Further, the billing information in the administrative data also contains the provider information (i.e., physicians and hospitals who have provided the treatment). When this data is viewed from the perspective of a specific treatment or procedure, the group of physicians who were involved in providing a particular type of treatment can be discovered. This information can then be used to understand the collaboration pattern among the physicians within and across hospitals.

The preceding discussion on the potential of the administrative dataset has primarily motivated us to pursue this research in order to explore how much we can learn about disease progression from a given dataset. For the scope and practicalities of the research, we have only focused on one particular chronic disease (i.e., type 2 diabetes), and have excluded from consideration some of the available information in the dataset, such as billing details and provider information. One feature of these datasets that inspired their use in this research is their untapped potential: there has been very little research done on disease risk prediction using administrative data to date. This is partly because such datasets have not been accessible to researchers before. Even though the administrative datasets were generated by individual providers, policy and related issues have created barriers to mining these data. A brief discussion of these issues is presented in the following section.

1.2.3 Policy-level changes in the healthcare industry

In the last two decades, several policy-level changes and their subsequent implementations have changed the ways in which administrative health data are recorded, formatted, stored and transmitted. With the aid of modern computer and network technology, these policy-level implementations enabled the data to reach the
research community and made them suitable for mining. This is also one of the reasons why it recently became possible to do this research in the Australian health context.

Two of the major issues in health data analytics have been standardisation and the availability of the data itself. Especially if the population of the dataset represents people from different geographical locations and socioeconomic conditions, it is important to be aware of the most relevant variables to be able to conduct a proper analysis. To understand the health status of the population, and patterns such as disease progression, it is crucial that all healthcare facilities (e.g., hospitals) record and report the data using a common format. This is a significant challenge even for most developed countries, as it demands multi-disciplinary effort, technical and medical expertise, resource allocation and training. Further, on an international level, countries follow different protocols and coding standards, making it difficult to utilise or merge datasets from different countries. Fortunately, the research community and international organisations have been working on solutions to this problem of standardisation. In the last decade, several milestones were achieved in standardising disease codes: for example, the ICD codes have been proposed and adopted by many countries, creating a common platform for analysing data from a wide range of sources. At present, for the Australian context, hospitals now follow the implementation of the ICD-10-AM coding format (see Section 3.3.1 for details) to record the health conditions and diseases of the patient throughout the country. Also, there are now established protocols on how healthcare data should be transmitted to governments, as well as to the research community. Further, policies have been implemented to ensure the privacy and security of the data as it is transferred across different bodies. All of these policies that have been formulated and implemented in recent years have motivated and allowed us to undertake this research.

Another factor that contributed to the research plan is the modernisation of the healthcare industry, especially of private health insurance companies. In fact, our research data comes from a group of private health funds i.e., organisation providing private health insurance. The complex process of recording, transmitting and obtaining this large amount
of health data would not have been possible if the government-established policy e.g., 2008 and 2010 health reform (Australian Department of Health and Ageing, 2011) had not recently been formulated and implemented. Newer policy has also opened up new frontiers of economic prospects for private health funds to provide added cover to citizens by providing quality healthcare. The economic incentive, policy and market competition in turn has boosted their efforts to minimise their costs. As discussed previously, a significant part of the healthcare burden, both economically and medically, comes from chronic diseases. So, naturally, the health industry is focusing more on identifying the high-risk chronic disease patients to adopt a suitable policy for them to minimise cost and improve quality of service. These policies have encouraged the industry to open up to researchers by sharing their vast stores of administrative data for investigation from data mining and analytical perspectives. This research is the direct outcome of such efforts, and is one of the earliest projects of its kind in this respect.

1.3 Key Research Questions

This thesis explores the use of administrative health data in understanding chronic disease progression, and subsequently goes on to develop a predictive model for identifying high-risk patients. In the process, we utilise several network-based methods and socio-behavioural parameters. The research questions that we aim to answer throughout the thesis can be stated as follows:

1. How can we utilise administrative data to understand chronic disease pathways?
   a. How can we best organise and filter administrative data for effective knowledge discovery, especially when the data is gathered from heterogeneous sources over a long period time?
   b. Do chronic and non-chronic patients follow different health trajectories?
   c. If they follow different trajectories, then what significant differences do they normally have?
d. Can we interpret the differences in health trajectories of chronic and non-chronic patients in terms of a baseline network?

2. Is it possible to make reasonable predictions of chronic disease risk from the understanding gained from the previous research question?
   a. Can we utilise graph theory and social network measures to compare the health trajectory of test (i.e., as yet non-chronic) patients with that of chronic disease patients (i.e., as represented by a baseline network)?
   b. To what extent do behavioural (e.g., smoking etc.) and demographic (e.g., age, sex etc.) risk factors contribute to chronic diseases?
   c. Overall, how can we construct a generic and adaptive predictive framework to assess any chronic disease risk based on the administrative data?

1.4 Approach

The research presented in this thesis is largely explorative, aiming to explore the potentials of the administrative dataset to answer the research questions stated above. Considering the context and features of the healthcare dataset, we selected type 2 diabetes as the chronic disease that would be the focus of the investigation. Overall, the research is presented in terms of a generic framework that can be implemented for other chronic diseases. In the thesis, the framework takes a longitudinal view of a diabetic patient’s health data, and analyses the pattern of the health trajectory to understand the influencers and inhibitors that contribute to health outcomes (i.e., whether or not the patient becomes diabetic). The differences between diabetic and non-diabetic patients’ admission histories are used to develop a baseline network. We consider the diagnosis codes that are present in the individual admission histories and formatted in ICD-10-AM as the main data elements of the research. We also use dates of admission as longitudinal information to represent the health trajectory of the patient.

In relation to the second set of research questions, we formulated several methods derived from graph theory and social network measures to compare the baseline network
against the test patients’ health trajectories. These methods give scores against several graph matching scores. We also derive several other scores from behavioural and demographic risk factors. Finally, the framework utilises several data mining methods to develop predictive models based on these risk factors. As mentioned previously, the present research and data analysis is explorative; we do not have prior information on which combination of methods and risk factors yields the most accurate prediction of diabetes risk. Therefore, we implement different combinations of methods and, based on the empirical result, decide on the optimum methods. This discussion, including why the optimum methods outperform the others, is presented in Chapter 5.

1.5 Contribution of the Study

This thesis explores a research problem that covers different fields, primarily - data science, network science and healthcare. To address the research problem, we develop a framework that has a theoretical and methodological contribution towards its related fields. Also, the outcome of the research can be potentially implemented into a predictive tool that can be used by different stakeholders in the healthcare industry. Considering these, we describe below the possible contribution of this thesis in three different contexts.

1.5.1 Theoretical contribution

This thesis proposes a framework that utilises graph theory and social network measures to predict the likelihood of chronic diseases developing in a patient. The type of healthcare data that we are using in this research is administrative data, which has rarely been used in previous research on chronic disease risk prediction. Therefore, this research is likely to contribute to the scientific community by showing the potential for the use of administrative data in chronic risk prediction.

The theoretical contribution of the research is likely to come from the methods and network based measures that we have proposed in the framework. These methods and measures can give insights about chronic disease risk and progression from health data.
For example, in the first part of the framework—‘understanding the progression of chronic disease using administrative data’—we have proposed a network formulation to representation of chronic disease progression. The framework shows that the diseases that occur as comorbidities of chronic disease (i.e., diabetes) have a networked relation that differs from the network of diseases of non-chronic patients. In the second part of the framework, we have proposed mathematical formulation of several new scoring measures (i.e., risk factors) derived from network theory and social network analysis (SNA). These new measures quantify the risk of chronic disease by looking at network-based similarities between a test patient’s health trajectory and the baseline network of the chronic disease. The implementation of the framework in the type 2 diabetes context has shown that some of these network-based measures are significant in predicting the risk for test patients. Therefore, the framework as a whole uniquely shows the networked structure of chronic disease comorbidities and the potential to use this structure for prediction using graph theory. These sets of concepts that the research has revealed are the primary theoretical contributions from the research.

1.5.2 Methodological contribution

The present research framework is more inclined to exploratory and analytical methods than to generating theories. Therefore, the significant contribution of this research will be its methodological contribution. The full methodological framework is divided into two steps, each of which introduces new methods required for analysis. For some parts, the framework proposes alternate methods (i.e., more than one), which are run in turns as part of the exploratory analysis. Later on, the prediction performances of these different combinations of methods are compared to understand which framework workflow, in terms of choice of methods, yields the maximum accuracy. This design is adopted to keep the framework as flexible and generic as possible, because administrative datasets may differ significantly in their properties based on the source organisation and country. Also, the context relevant to the chronic disease under investigation may vary. Therefore, the framework and underlying components should be adaptable to different contexts.
1.5.3 Knowledge translation for the health policy maker

The outcomes of this research should be of interest to several stakeholder groups connected to the healthcare industry. The research likely will offer little to physicians in terms of clinical diagnosis, as the input data for this research are generated once the physicians have completed the diagnostic procedures and evaluated the health condition. Rather, the contribution of this research is more targeted at policy makers working in government health organisations such as the AIHW, international organisations such as the WHO, or in public or private health funds.

The outcomes of the research and the knowledge generated from them can be directly translated into creating suitable risk assessment tools for the healthcare industries. One of the major aims of stakeholders and policy makers in these industries is to formulate health policies that can deliver the best possible health outcomes in addressing the burden of chronic disease and an ageing population. This aim is aligned with our research aim of understanding chronic disease progression. Accordingly, the network analysis methods presented in this research for analysing and visualising administrative health data have the potential to facilitate evidence-based decision making for stakeholders in the healthcare sector. The analytics can reveal trends among various comorbidities associated with a particular chronic disease for a targeted population (e.g., members of a health fund); for example, they can show which diseases occur more commonly with type 2 diabetes and their relative prevalence, and they can reveal patterns of disease progression that can help the analyst to forecast the burden of diseases that may ensue in the near future. These insights should help tremendously in formulating policies for resource allocation and budget planning. The first part of our research framework—‘understanding the health trajectory of chronic disease patients’—is focused on revealing these insights.

The most important and directly translatable contribution for policy makers should come from the second part of our research framework. This part of the framework is aimed at predicting the likelihood of chronic disease for a given patient. As per our direct
discussions with the private and public health funds and other interest groups before undertaking the research and receiving administrative data, it is evident that policy makers always face challenges in tailoring effective policy to different populations based on age, vulnerability, demographic condition, health status and other factors. As they formulate policies for the delivery of cost-effective and high quality care, they face several conflicting goals that need to be balanced. For example, to deliver high quality healthcare to consumers, they must allocate sufficient financial, medical and human resources and services. These resources and services are also scarce and exhaustive as the development and supply of health workforce and resources takes a lot of time, budget and pre-planning. Thus, it is a big challenge to plan, supply and distribute limited health resources in a comprehensive, efficient and transparent manner (Armstrong, et al., 2007). Further, market competition and financial models play a role in minimising resource allocation while achieving a balance between ensuring maximum turn-over for the stakeholders and providing sufficient quality of care to the consumers. One potential way to ensure this is to identify high-risk chronic disease patients in advance and plan preventive health management strategies for those high-risk groups. These can potentially reduce the future cost of accessing providers and improve quality of life for the patients, thus achieving a win-win situation for all. This is exactly the aim of the second part of our research.

1.6 Thesis Structure

This thesis is presented in six chapters, including this introduction chapter. In the following chapter, i.e., in Chapter 2, we review the existing literature to discuss theories and methods that are being used in understanding chronic disease progression and prediction of the disease risks. Because of the multi-disciplinary nature of this research, we begin with a brief introduction to prediction methods and evolution in healthcare, and discuss how computer technology has permeated the field, giving rise to the concept of electronic healthcare. We then discuss different groups of methods and their implementations in disease-related predictions, along with example models or systems built upon those methods. Following the review of existing literature, we discuss the
limitations of current methods and approaches. We then discuss the opportunities of improvements and applicability of newer methods in this area, followed by an overview of important social network theories that might be used in deriving the framework. Therefore, by the end of this chapter, the reader should have a clear idea of the theoretical and methodological background on which the thesis is constructed.

**Chapter 3** then introduces the context and conceptual background of the research. As we implement our framework on a real-world health dataset, it is essential to understand the dataflow and the different components of the data that are used during the implementation and analysis. Health datasets, particularly administrative datasets, vary significantly based on their location and the legislation, policy and technology that are used throughout the data flow process. As mentioned earlier, our context is that of the Australian health system, and our administrative data were received from private health funds. Chapter 3 discusses the specific features of our health dataset that are relevant for the implementation of the framework. This chapter also gives conceptual background on the core topics related to the framework.

Next, in **Chapter 4** we introduce our methodology and framework in detail, including its different methods and components. We begin by providing an overview of the framework as a whole. After that, we introduce several conceptual definitions that form the basis of the framework. Then we explain the methods involved with each of the components, as well as the relations between them. We introduce the related mathematics, equations and algorithms involved in the methods when necessary.

In the following chapter, i.e., in **Chapter 5**, we present the results and discussion. We describe the implementation of the proposed framework on an administrative dataset generated within the Australian healthcare context. As mentioned, type 2 diabetes was chosen as the chronic disease in focus. The results and findings of the implementation are presented, along with discussions alongside the relevant results. It is worthwhile to note that we intended to present the discussion of all results as a separate chapter. However, the methods involved in the framework are quite complex and the corresponding results
require lengthy descriptions. Therefore, separating the discussions of different framework components from their empirical result would make the discussions more difficult to comprehend. Hence, we present the discussion alongside the results and findings in an interleaved fashion that is easier to understand. This has made the chapter a bit lengthier. At the end of the chapter, we provide the summary and key findings from the implementation.

Finally, in Chapter 6, we present the conclusion. We provide an overall summary of the thesis followed by the implications of this research in terms of future research and practical applications. Following that, we discuss in more detail some directions for future research that are aligned with the current research topic. We conclude the chapter by discussing the limitations that were present in the implementation of the research framework.

After the chapters, we include a detailed Appendix section that might be of interest to readers. We have planned and presented this research with the aim that anyone with a moderate knowledge in healthcare and computer science should be able to replicate our research framework. Therefore, there are many topics that are not directly relevant to our analysis, but important to note for understanding the implementation. The Appendix section provides supplementary illustrations of this nature, including code snippets, definitions, data structures, software interface snapshots, mathematical notations etc. When necessary, we have included references to the Appendix within the main thesis.
Chapter 2: Background and Related Research

‘It appears to me a most excellent thing for the physician to cultivate Prognosis; for by foreseeing and foretelling, in the presence of the sick, the present, the past, and the future, and explaining the omissions which patients have been guilty of, he will be the more readily believed to be acquainted with the circumstances of the sick; so that men will have confidence to intrust themselves to such a physician. And he will manage the cure best who has foreseen what is to happen from the present state of matters...’

Hippocrates, *The Book of Prognostics*, 400 B.C.E.

In this chapter, we review the literature and discuss the theories and methods that are aligned with the research. Our research has a multi-disciplinary scope, as it involves topics from two broad domains: computer science and healthcare. The underlying theory and methods used in this research largely come from graph theory and network analysis, whose logics are strongly embedded in computer science. Another part of the research methods is derived from SNA. Although the theory of SNA originated in the social sciences, it has been adapted and used extensively within the computer science domain in recent years. For the present research context, we have further applied these methods in the healthcare domain—specifically on hospital admission data collected for administrative purposes. Therefore, our research utilises methods grounded in computer science, applying them in healthcare context; in a broad sense, this work can be categorised as belonging to the field of healthcare informatics.

The healthcare industry involves work from diverse fields and disciplines. Today, almost all of its subdomains have adopted some forms of information technology. From the management of resources to clinical decision making, healthcare informatics has a wide array of applications, each backed by large amounts of research. Since the scope of the present research falls within the sub-speciality of disease risk prediction using healthcare
data, in this chapter, we identify this scope within the broad discipline of healthcare informatics and review the major work done to date related to our method of predicting the risk of chronic disease.

Figure 2.1: Organisation of literature review concepts
2.1 Evolution of Healthcare Informatics

Since ancient times, illness, diseases and epidemics have shaped the course of human civilisation, including its politics, commerce and culture. Smallpox likely disfigured and killed Rameses V (the fourth Pharaoh) in 1157 BCE (Nelson and Williams, 2013, Ruffer and Ferguson, 1911). Bubonic plague and its coinfections contributed to the collapse of the Han empire in 160 CE and its subsequent hit halved the northern Chinese population during the third and fourth centuries (McMichael, 2001). Meanwhile, the Antonine Plague of 165–180 CE contributed to the fall of the Roman Empire (Fears, 2004) by devastating it on the western side of Eurasia. Until medical science developed to understand the underlying pathophysiology of these diseases, waves of epidemics continued to devastate various empires and parts of the world until the twentieth century.

Living at this current point in history, we are now facing tremendous opportunity and new health challenges. In terms of opportunities, we have modern medicine capable of treating or preventing various fatal diseases. We have highly developed computer, information and database technology aiding all subdomains of healthcare. People no longer die in significant proportions from smallpox, polio, measles and cholera-like epidemics. Conversely, we have different sets of diseases on the rise: chronic diseases in particular are greatly affecting the quality of life of our ageing population. We have huge amounts of structured data automatically generated within the healthcare system that is not mined, yet has the potential to yield deeper understanding of quality of care matrices. To illuminate the scope of the present research of utilising healthcare data to predict risk of chronic diseases, it is worthwhile to review briefly the evolution of healthcare informatics to understand how the field has reached its present state.

2.1.1 Prognosis: from ancient times to Greek medicine

Primitive medicine was essentially dominated by soothsayers, priests, oracles and various magico-religious authorities (Venot, et al., 2013, p. 142). It was the Greeks who established the foundation of modern medicine, with the notable works of Hippocrates,
Aristotle and Galen appearing within an approximate span of 800 years (600 BCE–200 CE) (Singer and Underwood, 1962, p. 1). Hippocrates is credited as the first person to consider medicine as a science, believing that diseases neither come down by God’s wrath nor by evil magic. Rather, he held the more rational view that they are caused by various environmental and lifestyle factors (Hippocrates, 400 BCE). Hippocratic medicine and philosophy tried to understand patient health and mental wellbeing, favouring passive or more natural methods of treatment (Kleisiaris, et al., 2014), reflected in the ethics of the Hippocratic oath (Edelstein, 2000). Related to our research scope, a key area of Hippocratic philosophy was ‘prognosis’ (Schiefsky, 2005, p. 17), which emphasises predicting the future state (progression) of the disease based on the present condition and clinical evidence. The Aphorisms, one of the most famous books in the Hippocratic Corpus (Hippocrates, 400 BCE), lists a series of brief generalisations on prognosis and clinical knowledge. Some notable extracts are:

- ‘When sleep puts an end to delirium, it is a good sign.’
- ‘Weariness without cause indicates disease.’
- ‘The old have fewer illnesses than the young, but if any become chronic with them they generally carry it with them to the grave.’
- ‘Those naturally very fat are more liable to sudden death than the thin.’
- ‘Convulsion supervening on a wound is deadly.’ (This is referring to tetanus.)
- ‘Those attacked by tetanus either die within four days or if they get through these they recover.’
- ‘It is fatal for a woman in pregnancy to be attacked by one of the acute diseases.’

Critical examination of the few examples above, as well as many others in the original corpus, demonstrates the clinical achievement of Hippocratic medicine in establishing sound guidelines for case description, clinical reasoning and observation, risk assessment, patient care and confidentiality, unparalleled in an age shrouded in medical follies, animism and superstitions.
To make a sound clinical decision involving risk assessment and disease prediction, one has to identify clearly the underlying factors and their interrelationships with potential to affect the progression of the disease. Although Greek medicine was arguably ahead of its time in establishing guidelines for finding those underlying factors for diagnosis and prognosis, it is not surprising that it lacked understanding in numerous areas on the pathophysiology of diseases. Different schools of thought existed within Greek medicine (e.g., the Koan and Knidian), often giving contradictory views. Knowledge of human anatomy was limited, as human dissection was taboo (Klaver, 2012). Many Greek medical beliefs were propagated through later civilisations and ultimately dismissed by modern medicine, such as the association of humoral pathology, involving four ‘humours’ in the body, with personality traits.

To sum up, the field of risk prediction of diseases, which forms the focus of the present thesis, can be traced back to Greek medicine, beginning with the concept of prognosis and establishing the bases of modern healthcare in principle. However, the accuracy and correctness of Greek practices are arguably weak compared to modern healthcare. From the time of the Greeks up until the present, history has seen the reign of several empires, through which and medical knowledge has flourished and been transmitted. Some notable periods were marked by the undivided Roman Empire, its division, the rise of Islam in the Arab Peninsula, Arabic medicine and the subsequent medieval awakening of Europe. An interesting thing to notice here is the lack of uniformity and access to healthcare knowledge during transitions between empires. A significant barrier in the dissemination of knowledge was perhaps language. For example, the Greek medicinal works known as the Hippocratic Corpus were written in Greek, while the Roman Empire mostly used Latin and the works of Islamic medicine were in Arabic. The lack of translation may have hindered the progress of knowledge.

2.1.2 Romans, Arabs and medieval Europe

After the Greeks, the Roman Empire contributed to medicine notably in the fields of public health, sanitation and hygiene (Singer and Underwood, 1962, p. 77). Hospital systems
were greatly organised and advanced by the Romans. Some areas that were improved are public hospital systems (Singer and Underwood, 1962, p. 46), resource management and codes of conduct. The western part of Roman Empire collapsed around 476 CE. However, the Eastern Roman Empire, which existed until it fell to the Ottoman Turks in 1453, preserved much of the Greek medical works (Singer and Underwood, 1962, p. 66). In fact, much of the original Greek corpus was translated by Romans into the Syriac language.

One striking argument for the importance of maintaining a universal language (Venot, et al., 2013, p. 11) for defining diseases (e.g., ICD) and protocols can be understood from the fact that different civilisations used different languages. Therefore, to properly understand the works of previous generations, one must be capable of reading their books or manuscripts without significant loss in translation. Assimilating knowledge from texts written hundreds of years previously in a language not known to many contemporary people seems extremely challenging, if not impossible. Still, this feat was achieved by the wonderful people belonging to the Roman or Arab empires or to the Spanish peninsula. As stated earlier, a number of Greek works were translated into Syriac by Romans. During the Islamic Empire in the Arab Peninsula, many of the original and translated works of the Romans were further translated into and studied in Arabic. Arabic science also progressed significantly in terms of its fundamental contributions in medicine, mathematics (e.g., algebra, Arabic numerals) and astrophysics (Rashed, 2013). In fact, many modern mathematical and prediction models are indebted to Arabic arithmetic and algebra. As the Spanish peninsula was inundated by the Arabs as early as the eighth century, a bilingual generation began to appear. They were able to further translate the Arabic texts into Latin. Christian Europe, or the West, finally obtained access to this body of knowledge via the Arabic-Latino literature. From the thirteenth century onwards, a large number of universities were established across Europe (Singer and Underwood, 1962, pp. 70–72), and the practice of medical science and related topics became much more oriented around them.
2.1.3 Renaissance and the industrial revolution

The Renaissance in Europe spanned between the fourteenth and seventeenth centuries. Having begun as a cultural movement in Florence, it soon revolutionised the fields of science and medicine. Several notable events took place to assist this revolution: the printing press was invented, making available printed copies of improved translations of Greek works. The invention of the microscope made it possible to look into the minute structures of animal bodies. Knowledge of human anatomy and physiology were greatly enhanced, rendering many previous medical beliefs outdated or invalid. This indeed helped the greater scientific community to understand the origins and prognosis of diseases better than ever.

Following this period, the industrial revolution (approximately 1760–1840) in Europe came, which resulted in rapid urbanisation, economic growth and lifestyle changes. The improved living standards, especially in urban or industrial areas, along with the contemporary agricultural revolution, brought positive changes in public health and nutrition. For example, hospitals had more hygienic guidelines, the prevalence of various diseases (e.g., malaria, rickets, scurvy) declined rapidly (Singer and Underwood, 1962, p. 181), and life expectancy increased. One of the biggest contributions in medicine, the germ theory of diseases, matured during and immediately after this period. The germ theory scientifically explained the pathophysiology of different communicable diseases such as smallpox, plague, puerperal fever, cholera and anthrax.

The decline of the plague and improved life expectancy brought attention to non-communicable chronic diseases. Factors like industrial and agricultural revolution, urbanisation and changed lifestyle contributed to the higher prevalence of obesity, which is the common risk factor for multiple chronic diseases.

A significant part of public health, clinical decision making and medical prediction relies on basic statistics. During the industrial revolution, a substantial amount of novel work was done in this field. In particular among the probability theories, Bayesian statistics (Bayes
and Price, 1763) was proposed by Thomas Bayes and presented posthumously. Others notable work included that of Laplace, who, apart from contributing to mathematics, astrophysics and statistics, popularised the notion of Bayesian probability (Stigler, 1986, pp. 97–98, 131). However, the true potential of Bayesian probability or statistical inference (Venot, et al., 2013, p. 6) went unexplored in the medical literature until the twentieth century, when the medical domains matured, and particularly began to integrate digital computers into their research.

2.1.4 Introduction of computer and medical informatics

The invention of computers has fundamentally influenced medical science and all of its subdomains. Although the idea of the modern programmable computer can be traced back to the early nineteenth century (Halacy, 1970), working versions of digital computers began to be developed from the 1940s. At that time, these were not yet practical to use in the medical sector for record keeping or diagnosis purposes, yet several advancements were made subsequently in computer technology that finally overcame the usability issue in the medical domain. We could mention five such breakthroughs of particular relevance. The first was the invention of the transistor in 1947 (for which the inventors John Bardeen, Walter Brattain, and William Shockley were awarded Nobel Prize). Transistors quickly replaced bulky vacuum tubes and made computers smaller, cheaper and more energy-efficient. The second was the invention of the microprocessor. First becoming available commercially around 1971, microprocessors paved the way for general purpose personal computers. Third, the development of application software, both customised and generic (word processing, spreadsheet), greatly helped to replace paperwork with computer-based records. Fourth, databases helped to popularise the idea of analysing healthcare data, as they facilitate saving and organising data and offer efficient retrieval methods. Last but not least was the advent of the internet around the late 1990s; local networking, database systems and client–server software architecture all synergistically contributed in shifting paper-based medical records onto computer-based systems.
Similarly, the internet made it possible to share and gather knowledge over large distances quickly and conveniently.

Figure 2.2: Evolution of healthcare informatics and related fields with computer technology
These five key developments allowed medical informatics to evolve gradually into a scientific discipline (Venot, et al., 2013, p. 2). In fact, it is possible to trace this development by making a crude comparison of the frequency of appearance of the terms ‘medical informatics’, ‘health informatics’, ‘biomedical informatics’ and ‘healthcare informatics’ in the literature over time (Michel, et al., 2011). This evolution is shown in the Figure 2.2. Interestingly, we can see how the frequencies of these terms leapt soon after breakthroughs in computer technology, thus aligning the developing science with the shifting paradigms of technological advancement.

2.1.5 Electronic health records and administrative data

As part of social security measures by governments, national health insurance systems started to become common after the 1950s. For example, in the US, Medicare was founded in 1956 (Robinson, 1957), while Australia’s Medibank was initiated in 1975, later reformed as Medicare in 1984 (Kewley, 1973). Eventually, in order to formulate better policy to ensure optimum value for money in healthcare, there was an increased need to understand the healthcare inputs for different stakeholders. These healthcare stakeholders related to policy formulation and research can include:

1. Government organisations (e.g., health department, bureau of statistics etc.);
2. State or local area government bodies (e.g., state health departments, when applicable);
3. Private health insurance providers;
4. Healthcare providers (e.g., hospitals);
5. Independent research institutions (e.g., universities).

Healthcare providers routinely collect data for their internal records, as well as for reporting to different governing bodies. These records are commonly known as administrative data, as they are used for administrative purposes, not for any clinical reasons. Initially, these records were completely paper-based. When personal computing became sufficiently affordable and powerful, healthcare providers began to shift to
electronic means of keeping records. However, in most cases, data was still collected in paper format, as it was more convenient, and policy makers did not yet see the potential use of electronic data stores. Concurrently with the popularisation of technologies such as the World Wide Web, email and database systems after the 1990s, this potential for keeping electronic administrative records began to be realised. The concept of e-health records also became more widely recognised, where the complete medical information of the patient is to be stored in a central database. This data can be accessed by patients, general physicians or emergency medical teams with convenience to provide faster and more accurate treatment.

After the 2000s, the concepts of big data and data mining also gained widespread recognition and, accordingly, administrative data was recognised for its research potential. Unlike data from clinical trials, which are run on smaller populations, administrative data offer a representation of the health conditions of the larger population that can be made easily accessible to researchers. This offers ample opportunity to understand the prevalence of diseases by doing epidemiological research (Warren, et al., 2002). Further, healthcare data contains time information, making it suitable for longitudinal studies and understanding patterns of disease occurrence. This has the potential to offer ways of predicting future stages of health conditions of an individual or a particular cohort (e.g., the ageing population). The focus of this thesis lies in this field—to develop a predictive framework for identifying chronic disease risk for patients.

2.2 Statistical Methods of Disease Risk Prediction

Different statistical methods are used extensively in clinical decision making. For any disease risk prediction, whether by a computer or by a physician, one has to (1) understand the factors or variables that are responsible for the disease, and (2) based on the presence or absence of the factor, assess the patient’s risk of disease. A similar process is used in clinical diagnosis. Now, to make these clinical decisions of diagnosis and risk prediction, one has to apply logic or deductive reasoning (Ledley and Lusted, 1959).
This natural reasoning process in medical research essentially is to examine a sufficiently large number of patients having the same disease (cohort) to check the common variables within the cohort. Then, if these variables are absent from another cohort of patients who do not have the disease, the variables are held to be connected with the disease. The variables could be symptoms (e.g., diagnostic test results, clinical observations) or physical characteristics (e.g., age, sex). However, in most cases, a set of responsible underlying variables cannot be considered as exclusively belonging to the diseased patients. For example, imagine a scenario where a cohort of obese people is found to have more heart disease prevalence than a non-obese cohort. Now, in most real study data, there should be some degree of exception. Not all obese people will have heart disease, and similarly, some non-obese people will have heart disease. Therefore, the variables will not be assigned a categorical risk status (i.e., either they are responsible or they are not); rather, they are assigned a contributing probability of risk. Based on the found likelihood of heart disease being associated with obese people, one should carefully examine whether or not the association is significant. This scenario is very common in medical research, and statistical analysis is particularly suitable for this type of study. Statistical methods are often used as an integral part of other methods as well as to present the result numerically (e.g., p-values) or graphically (e.g., histogram).

In different branches of epidemiology, comparison studies and other clinical trials statistical methods are widely used. For example, descriptive statistics are used to understand the relevant variables and their distribution: hypothesis tests, t-tests and analysis of variance (ANOVA) or significance tests are widely used to compare two different cohorts; risk statistics particularly focus on assessing the odds of contracting a disease by people exposed to certain factors compared to others; and correlation statistics deal with describing how closely two variables are related (Barton and Peat, 2014).

Regression models are often used in predicting the future state of a variable (outcome variable) given one or more variables (explanatory variable). The outcome variable can
denote the risk of disease when applicable. Regression models are often used in predictive modelling (Barton and Peat, 2014, p. 205): for example, age and gender (explanatory variables) can be used to predict normal values in Body Mass Index (BMI). Regression models can also be used for hypothesis testing by examining the effect of an explanatory variable on an outcome variable after adjusting for other important explanatory factors. For example, the model can use age and gender to predict BMI, and in turn be used to test the hypothesis that groups with different exercise regimes have different BMI values.

Statistical methods have several limitations; it is sometimes said satirically that ‘statistics can prove anything’. The advantage and accuracy of any risk prediction model highly depends on its methods and design, including the identification and choice of variables, appropriate statistical methods and research questions. Correlation tests can lead researchers in the wrong direction if not handled properly. Two variables can be shown to be related, but this does not readily prove that one is causing the other; that is, correlation is not causation. A well-designed statistical method should investigate the underlying cause of the correlation once it is shown by the test. In some cases, correlation can show false negative results, as in the case of non-linear relationships. Also, if there are any missing variables in the experiment, statistical methods can give erroneous results, although this is a limitation for other methods as well.

2.3 Data Mining-Based Disease Prediction

Traditionally, clinical decisions—including diagnosis, treatment and disease prediction—are made by doctors’ intuition, knowledge and experience with the aid of various clinical and diagnostic tests. This practice often leads to unwanted biases, errors and excessive medical costs (Palaniappan and Awang, 2008). Clinical decision-making processes can be supported by computer-based patient records, and this has been shown to improve several performance matrices, such as improving patient safety and outcome, and reducing medical errors and unwanted practice variation (Wu, et al., 2001). Computer-based information management, including, for instance, patient record keeping and database management, prove to be cost-effective and superior in performance compared
to traditional record keeping. Computer-based information management has been
popularised as an idea and begun to be integrated; in the process, computerised data
keeping systems have generated huge amounts of data. This not only includes patient
health records, but also includes information on patient–doctor interactions, insurance
claim data, medical history and referral information. However, this wealth of data often
goes untapped. Recently, data mining methods have been devised in search of ways to
utilise these data.

2.3.1 Score-based models

Scoring-based methods are perhaps one of the earliest and most intuitive methods of
disease prediction. In traditional medicine, before the aid of computerised systems,
doctors were obliged to observe the symptoms and medical history of patients and to
evaluate them based on a combination of experience and well-established case studies to
predict the risk of disease or comorbidity. This idea of symptoms-based prognosis has led
to the development of many scoring-based methods to ease and standardise the
assessment of disease and related risk in various healthcare settings. In these methods,
scores are assigned to various factors such as physiologically observable conditions,
demographic information or family history. Once a patient’s score is calculated, it is then
normally evaluated against an interpretation table that describes the probable range of
scores and their corresponding meanings. The ‘risk score’ thresholds are established
based on clinical studies of different cohorts, often employing multivariate regression
analysis methods. Risk score-based methods are simple to use, take less time to assess
and are often deployed in web-based health assessment calculators for consumer use.
The following Table 2.1 shows some sample factors that are considered while assigning
scores.
### Table 2.1: Sample parameters to score against in score-based system

<table>
<thead>
<tr>
<th>Category</th>
<th>Example of risk factors for scoring</th>
</tr>
</thead>
<tbody>
<tr>
<td>Demographic factors</td>
<td>Age, sex</td>
</tr>
<tr>
<td></td>
<td>Population subgroups based on age (e.g., children, adult, elderly), ethnicity (e.g., Asian, African)</td>
</tr>
<tr>
<td>Behavioural factors</td>
<td>Smoking, drinking, duration of physical activity, sun exposure time, etc.</td>
</tr>
<tr>
<td>Genetic</td>
<td>Family history of diseases</td>
</tr>
<tr>
<td>Biomedical factors</td>
<td>Physical traits: weight, height</td>
</tr>
<tr>
<td></td>
<td>Diagnostics: blood pressure, cholesterol, blood glucose level, etc.</td>
</tr>
</tbody>
</table>

Various score-based systems are currently in use. For example, the Charlson Comorbidity Index (Charlson, et al., 1987) was proposed as early as 1987, and predicts the 10-year mortality for a patient by ranking a range (total 22) of demographic factors (e.g., age) and comorbid conditions (e.g., heart disease, cancer, AIDS). A higher overall score means a higher chance of mortality of the patient within next 10 years. The Charlson Comorbidity Index has been extended and adapted into different variants, such as Charlson/Deyo, Charlson/Romano, Charlson/Manitoba, Charlson/D’Hoores, Charlson/Ghali or Charlson/Dartmouth. Some of these methods have migrated to newer coding schemes such as ICD-9. The Elixhauser index (Elixhauser, et al., 1998) is another similar index for finding mortality rates (discussed in details in the next chapter). It has slightly better prediction performance (Sharabiani, et al., 2012), especially when predicting mortality beyond 30 days. Another widely used scoring system is APACHE-II (Wong and Knaus, 1991) (versions I–IV are available, version IV is only used in the US). It scores against 12 variables with a range of 0–71 to assess the condition of ICU patients in the first 24 hours of admission (higher values indicate greater severity). Other similar ICU scoring methods include SAPS (Simplified Acute Physiology Score, version 1-3) and MPM (Mortality Prediction Model, version 1, 2). Later versions of these systems usually incorporate larger datasets, calibrated against different hospital settings, and can give predictions of different phases of ICU admission (Breslow and Badawi, 2012). These models provide a
good way for physicians to assess patient condition, often without the help of sophisticated software. They usually work well in a specific healthcare setting like ICU, where it is important to determine quickly the required aggressiveness of treatment based on the patient’s chance of mortality. However, these score-based systems are mostly unsuitable for predicting a patient’s long-term chance of contracting a chronic disease where a huge volume of admission and related medical history data is available.

Scoring-based methods are also used to group diseases into several types according to their complexity. This is often done to estimate the hospital resources needed to ensure quality of care and facilitate reimbursement. The Diagnosis-Related Group (DRG) (Fetter, et al., 1980) is one set of disease codes, implemented in the US in the early 1990s and later adopted internationally with local modifications in different countries. The original motivation of the DRG was to develop a classification system that could identify the ‘products’ received by the patient to aid in medical financing. However, as the healthcare industry has evolved since its initial introduction, the DRG codes have been modified and extended to support different objectives with a higher level of sophistication and precision (Baker, 2001). Modifications were also made by governments to meet national policies. A grouper software based on ICD codes scores the medical conditions of the patient, including diagnosis codes, comorbidities, discharge status and demographic information, and determines the DRG grouping, on which medical financing and reimbursement are planned. While this grouper software is useful for hospital resource management and cost planning, it does not predict future disease risk, but rather gives an estimation of the current condition. Overall, these scoring-based methods are limited to a smaller scope within the healthcare setting, and mostly cannot capture the interrelations or comorbidities of different disease codes. Table 2.2 shows some example of different scoring systems in practice.
Table 2.2: Example of different scoring systems

<table>
<thead>
<tr>
<th>Score Name</th>
<th>Predicts</th>
<th>Prediction Range</th>
</tr>
</thead>
<tbody>
<tr>
<td>Framingham Risk Score</td>
<td>Different cardiovascular diseases, type-2 diabetes</td>
<td>10 years (for most cardiovascular diseases)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>30 years (one variant for CVD)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>8 years (diabetes)</td>
</tr>
<tr>
<td>Reynolds Risk Score</td>
<td>Cardiovascular disease</td>
<td>10 years</td>
</tr>
<tr>
<td>SCORE</td>
<td>Cardiovascular disease</td>
<td>10 years</td>
</tr>
<tr>
<td>QRISK</td>
<td>Cardiovascular disease</td>
<td>10 years (QRisk, QRisk-2)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Lifetime (QRisk-lifetime)</td>
</tr>
<tr>
<td>QDiabetes (QDScore)</td>
<td>Type-2 diabetes</td>
<td>1–10 years</td>
</tr>
<tr>
<td>Diabetes Risk Calculator</td>
<td>Type-2 diabetes</td>
<td>8 years</td>
</tr>
<tr>
<td>ARIC Diabetes Risk</td>
<td>Type-2 diabetes</td>
<td>9 years</td>
</tr>
<tr>
<td>Calculator</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Gargano Mortality Risk</td>
<td>Mortality of type-2 diabetes pre-diagnosed patients</td>
<td>2 years</td>
</tr>
<tr>
<td>Score</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Charlson Comorbidity</td>
<td>Mortality</td>
<td>10 years</td>
</tr>
<tr>
<td>Index</td>
<td></td>
<td></td>
</tr>
<tr>
<td>APACHE-II</td>
<td>ICU mortality</td>
<td>24 hours of admission</td>
</tr>
<tr>
<td>SAPS II</td>
<td>ICU mortality</td>
<td>24 hours of admission</td>
</tr>
</tbody>
</table>

2.3.2 Collaborative filtering

Collaborative filtering is a set of data mining techniques generally used to filter or understand the collaboration or behaviour of users and to understand the pattern of the collaboration (Terveen and Hill, 2001). This broad idea has been scaled down to find application in many different fields of science. In our present research, collaborative filtering is related to predicting a user’s choice by analysing the choices (or collaboration) of similar users. It is often used as a recommender system by learning users’ past choices of items (Adomavicius and Tuzhilin, 2005). The basic proposition behind this is that, if a certain user group habitually chooses a set of items and another user chooses an item from that set, it is likely that the user will also like the other items of the set. The
preference information might be explicit (e.g., user ratings, likes on posts) or implicit (e.g., browsing history, bounce rate). As part of the recommender system, collaborative filtering is used extensively to track a user’s activity or choices in various online sites and to provide a more personalised experience, including suggestions and advertisements based on the choice profile. For example, a collaborative filtering-based recommender system is used to suggest items in e-commerce sites like Amazon (Linden, et al., 2003), movies in Netflix (Bell and Koren, 2007, Bennett and Lanning, 2007) or news (Billsus, et al., 2002), and even to find compatible matches in online dating services (Brozovsky and Petricek, 2007).

The methods of implementing collaborative filtering vary widely in the literature. The method of finding similarity can be loosely categorised as memory-based or model-based, or a hybrid of the two (Das, et al., 2007). In memory-based collaborative filtering, all user preferences for items are loaded or mapped into memory, and then similarities with a test user’s preferences are computed using certain methods. The test user’s preferences are predicted based on the items that have the highest weight of similarity. Alternatively, in many cases, a limited number of user preferences from ‘nearest neighbours’ are calculated (Davis, et al., 2010) by finding the similarity between two users and iterating the process over all users. The two most common similarity calculation methods are Pearson correlation (Resnick, et al., 1994) and vector similarity (Breese, et al., 1998). Memory-based collaborative filtering is relatively straightforward, intuitive and normally performs well. However, loading entire user preference databases for similarity calculation is often resource-intensive, especially if the data is sparse, which is true for most online items. Besides, as the item database or vector space is tightly structured, introducing new items is often difficult, as it requires reorganisation of the structure.

Model-based collaborative filtering methods generally utilise machine learning and data mining algorithms. There is an extensive variety of algorithms for implementing model-based filtering, which vary considerably in terms of performance and prediction accuracy (Si and Jin, 2003). Different clustering algorithms are often used in model-based
collaborative filtering. Here, users are clustered into different classes using the training dataset of user-item preference. The active user is then classified using the same algorithm to discover to which cluster it belongs, and choices are predicted from that cluster. Bayesian clustering and Bayesian network models (Breese, et al., 1998, Dempster, et al., 1977) are often used as the clustering algorithm. Another model-based collaborative filtering method is k-means clustering (Xue, et al., 2005). Here, the users are grouped into k clusters or classes by vector quantisation, where the users of the same clusters have the minimum mean. However, finding clusters with the minimum mean property is an iterative process and computationally expensive (NP-hard). Therefore, approximations using heuristic algorithms (Kanungo, et al., 2002) are often utilised to work around. Besides these, Markov decision process (Su and Khoshgoftaar, 2009), an extension of state-based Markov chain modelling, is also implemented in some collaborative filtering methods.

Unlike memory-based methods, model-based algorithms are usually scalable and faster, as they can efficiently handle the sparsity of the data when the dataset is too large. However, constructing the model is often expensive, and can make it difficult to introduce new data into the training dataset. Several algorithms have been proposed to increase performance by reducing complexity. For example, principal component analysis (Kim and Yum, 2005) is used to reduce the number of parameters; fuzzy clustering (Honda, et al., 2001) has been used to approximate the clustering process and missing values; and singular value decomposition has been used to reduce dimensionality (Sarwar, et al., 2000). Again, this introduces a trade-off between prediction performance and scalability. As the complexity of the model or data is reduced, scalability and performance increase, at the cost of prediction accuracy.

Collaborative filtering methods have great potential in predicting diseases because of their comorbid nature in patients; that is, many diseases or symptoms tend to occur simultaneously. However, only limited research has been done in this particular field (Hassan and Syed, 2010). Davis et al. presented ICARE (Davis, et al., 2008, Davis, et al.,
2010), a method based on clustering and collaborative filtering to predict risk for individual patients based on their own medical history and that of other similar patients. The ICARE system uses vector similarity as a collaborative filtering method, where ‘users’ are patients and ‘items’ are diseases. This method is extended by using the inverse frequency of diseases to capture the effect that having rare diseases in common has more impact than having a trivial disease in common. While this method incorporates several features, it still does not consider the sequence or timing of diseases, which is very important in disease behaviour. For example, the risk of disease C might be higher if the patient has disease A and then disease B in sequence. However, if the patient develops disease B before A, the chances of getting C might be low. Vector similarity does not capture this sequence. Further, gaps in time between occurrences of diseases are also important. These longitudinal aspects of disease occurrence are not captured in the ICARE system. While our research method does not directly employ collaborative filtering, our approach implicitly captures the essence of it. Further, we have taken into account the effects of sequence and time gaps between occurrences of diseases.

2.4 Networked Approach in Healthcare

In generic use, network means a group of people, objects or any other entities connected by some sort of relation, such as a road network connecting places, a friendship network connecting people, or a citation network connecting scholars with the relation ‘who-cites-whom’. In mathematics or computer science, a network holds a similar meaning, and can be represented as a graph. A graph has a set of nodes (also known as vertices or actors) and a set of edges (also known as ties). Nodes symbolise any entity and edges symbolise relations between entities. Problems from different scientific domains can be modelled as networks and visualised as a graph, and range of graph-theoretic methods and algorithms are found in the literature for analysing or solving problems.

In the context of disease prediction from healthcare data, many statistical and data mining methods do not take the relationship between diseases and symptoms into account explicitly. However, in reality, diseases—especially chronic or non-communicable
diseases—do not occur in isolation (Barabási, 2007). They often share a common risk factor, which can be genetic, environmental or behavioural. These risk factors have a synergistic effect (Burton, et al., 2007, Loscalzo, et al., 2007) on the health outcome that is often hard to predict if considered in isolation. For example, obesity is a common risk factor for many chronic diseases (Must, et al., 1999), including type 2 diabetes and heart disease. Simultaneously, obesity itself is the product of other risk factors related to lifestyle and behaviour. Also, the relations are not one-directional, and can often be part of a feedback loop. Thus the risk factors have a synergistic effect on one another. These relations are explicitly considered, studied and visualised in a networked approach.

A significant portion of network-based analysis is done on genetic associations of diseases. The human body has 23 pairs of chromosomes responsible for transferring genetic information as a blueprint from one generation to another. Chromosomes are made of DNA (deoxyribonucleic acid), which controls the production as well as production time of proteins, which in turn play a major role in determining the function and structure of cells, tissues and organs (World health Organization, 2014). Overall, a set of genes (i.e., genotype) in our DNA is responsible for a particular trait (e.g., eye colour). Further, the interaction between genotypes and the environment results in the physical expressions or characteristics of that trait commonly known as phenotypes. The overall relations between genotypes, environment and phenotypes are complicated.

Dysfunctional gene behaviour is commonly termed a mutation (World health Organization, 2014). These mutations are often responsible for causing illnesses. Moreover, if the gene mutations exist in the egg or sperm cells, children can inherit defective genes from their parents. As a consequence, researchers are trying to understand disease pathogenesis by mapping gene expression and the associated proteins that act on the same pathway (Hidalgo, et al., 2009). The networked approach is perhaps the most suitable and natural way to understand these relations, as it explicitly considers different micro- (e.g., DNA) and macro- (e.g., organ) level actors as nodes. These nodes are related in multiple ways; that is, they have ‘many-to-many’ relations from a network.
perspective. Numerous studies, especially in the biomedical domain, have employed a networked approach in understanding gene behaviour (Ideker and Sharan, 2008). Linkage analysis and association studies (Botstein and Risch, 2003) have been shown to be successful in identifying causative genes for many Mendelian or single-gene disorders (Chial, 2008). According to the Online Mendelian Inheritance in Man (Hamosh, et al., 2005), an extensive online database focused on inherited genetic diseases in humans, there are 3,552\(^2\) human genes of known sequence with a known phenotype-causing mutation, and 5,733 human phenotypes with a known molecular basis (OMIM, 1985).

In contrast to the Mendelian disorders, for diseases that are more complex (e.g., autism) or chronic (e.g., inflammatory bowel disease, diabetes, coronary heart diseases), a larger genomic region must be analysed to understand the genetic association. This can include tens or hundreds of candidate genes within the research scope, making the analysis a more difficult task. However, studies have suggested that genetic diseases are caused by functionally related genes (Oti and Brunner, 2007). Further network-based studies have shown that these causative genes for the same or similar diseases will generally reside in the same biological module, either a protein complex (Lage, et al., 2007), a pathway (Wood, et al., 2007) or a sub-network of protein interactions (Lim, et al., 2006).

Different statistical methods are also incorporated in network-based gene-network analysis. For example, Wu et al. (2008) proposed a regression-based generic tool called CIPHER to predict disease genes by prioritising candidate genes from the whole genome instead of genetic loci. This can perform genome-wide scans of causative genes for most of the recorded human phenotypes. Letrovsky and Kasif (2003) proposed a probabilistic binomial model combined with a Markov random field propagation algorithm. The model aimed at finding node neighbours in a protein–protein interaction network to predict which proteins are more likely to interact. Again, this model is based on an assumption similar to that discussed previously: network neighbours (i.e., proteins) are more likely to share functions than nodes that are distant. Other notable methods involve protein

\(^2\) As of 11 May, 2016.
function prediction using (a) logistic regression models (Lee, et al., 2006), (b) sequence similarity and interaction databases (Espadaler, et al., 2005), (c) message-passing algorithms or belief propagation (Leone and Pagnani, 2005) and (d) annotation of protein function across species by network comparison (Bandyopadhyay, et al., 2006). The following Figure 2.3 shows some examples of the network-based approach in disease or associated biomarker prediction.

![Figure 2.3: Sample network-based approach in disease or associated biomarker prediction](image)

2.5 Social Network Analysis in Medical Informatics

Social Network Analysis, commonly known as SNA, is based on network and graph theories. However, it stands out as its own domain because of its breadth of theoretical focus, encompassing the dynamics and structure of real-world networks with application in different domains, many of which have only recently begun to receive attention. As the name suggests, SNA is particularly aimed at understanding the formation and evolution of...
social ties (e.g., friendship, collaboration) between actors (e.g., people). In real-world scenarios, social ties between people do not form randomly and at the same time. For example, given a classroom, students who are unknown to each other do not readily befriend others. There are specific extrinsic and intrinsic factors that dynamically affect friendship choices, such as personality, friendliness, presence of mutual friends or academic performance. SNA methods and theories are particularly focused on addressing questions such as how social ties are formed between individuals, how they evolve over time, which actors have a more central role, which patterns of ties will yield better performance and which tie will form next (Wasserman, 1994).

The origin of the idea of ‘social networks’ can be traced back to the late 1800s, when Émile Durkheim and Ferdinand Tönnies did their research on social groups forming as a result of interaction between individuals (Aldous, et al., 1972). Subsequently, some notable research on social networks was conducted in the 1930s by different researchers from the fields of psychology, anthropology and mathematics. Some of the earliest experimental studies on SNA were done by Bavelas and Levitt (1950, 1949), Milgram (Small-world experiment, 1967) and Sampson (1968). The last of these, Sampson’s monastery experiment, can be credited as the first longitudinal social network experiment to systematically analyse change in the structure of a social network (e.g., friendship network) over time. Subsequent longitudinal studies on social networks can be credited to Newcomb (1961), Sanil et al. (1995), Snijders et al. (1990, 2010), Huisman and Snijders (2003) and McCulloh and Carley (2008). Doreian and Stokman (1997) produced a seminal text on the evolution of social networks. In their book they identified a minimum of 47 articles published on social networks that included some use of time, as of 1994.

Although SNA was originally intended for the social science domain, it has quickly spread to other domains, including medicine and public health. Because SNA is based on a solid theoretical background derived from graph theory, along with the advantages of its own theories, SNA is inherently strong in understanding collaboration patterns and subsequent network performance. As a result, it is widely used to understand collaboration between
physicians, physicians and patients, or across a hospital network. For example, Uddin et al. (2011, 2015) proposed an SNA framework for understanding the performance of collaboration (between physicians) and coordination (between hospitals). Using network centrality theories, they explored structures yielding optimal performance among different networks: the patient-centric care coordination network (PCCN), hospital-rehab coordination network (HRCN) and physician collaboration network (PCN). SNA is also used to understand research trends and map knowledge structures in the healthcare domain, for instance, in obesity research (Khan, et al., 2016, Uddin, et al., 2015).

Recent technological advancement has enabled hospitals to electronically collect and report healthcare data commonly known as administrative data (e.g., hospital admission, discharge reports). Along with concepts like ‘big data mining’, SNA-based approaches in administrative data mining have gained much attention among researchers (Baglioni, et al., 2013), as these datasets are inherently linked. For example, administrative data contains information about which patients are treated in which hospitals and whether they have been transferred between healthcare centres—for example, between public and private hospitals, between hospitals and rehabilitation centres, or towards a more specialised unit. Effective coordination for patient transfers between hospitals or related entities has been studied using administrative data and SNA (Anderson, 2002); some of these studies were mentioned earlier (Uddin, 2011, Uddin, et al., 2015). Apart from hospital information, administrative data contains a summary of diagnosis information. This has great potential for analysis on a large population level to understand the nature of comorbidities (i.e., which diseases occur together). SNA has been used previously in understanding the progression of chronic comorbidities (Luijks, et al., 2012) and predicting diseases (Folino, et al., 2010, Khan, et al., 2016, Khan, et al., 2017). However, the actual contexts in terms of healthcare settings, approaches and entities considered in SNA vary widely (Baglioni, et al., 2013, Chambers, et al., 2012).
2.6 Limitations of Existing Theories

The theories and methods of prediction in contemporary literature both have some limitations in the analysis of complex networks. These limitations vary across the different approaches, but in general, they tend to relate to the limited perspective of dynamic networks. Most of the theories focus on the underlying social phenomenon, such as interpersonal liking, information diffusion in certain communities, profit maximisation or ethnographic study, and make attempts to discuss and explain the underlying causes and their effect on the overall social structure. In most cases, the network under examination has little or no temporal components and the dataset is quite small—generally confined to a fixed community. In contrast, the healthcare datasets are quite big, having large numbers of different actors who often may have temporal and spatial attributes. On top of that, these networks are, in most cases, continuously evolving. To analyse the dynamics of such networks, we argue that present theories should be extended to address these issues.

Similarly, existing methods have attempted to explore the dynamics of longitudinal social networks. While they all have considerable strength in analysing certain aspects of healthcare data, they also have some assumptions and limitations regarding network size, structure and social theory. For example, Markovian models have the assumption of the memoryless property and do not consider exogenous changes in the network. Multi-agent simulation models are also bound to the underlying social theories. Some methods focus on analysing different properties of dynamics (e.g., Sampson’s approach on group formation) or trying to explain how different properties have changed (e.g., statistical methods) by correlating various factors from the observed network at different points in time. In cases where the networks are still changing, and the future goal is uncertain, present models cannot be applied efficiently to discover the underlying dynamics. In addition, in the case of a large healthcare network, it is often not possible to observe all the fine-grained changes for all the actors, as this would be computationally expensive.
Further, these models are quite impractical for the task of comparing and contrasting longitudinal healthcare networks under some benchmarks of centrality rankings.

Generic prediction models have their strength, in most cases, when the relation is simple and does not depend on many variables. Regression analysis, for example, might not work as well when we want to model complex phenomena, which rarely have simple explanations. For instance, in a healthcare dataset there may be multiple actors, including patients, physicians, specialists, hospitals and service providers. Each pair of actors has different links between them, resulting in a complex web of interaction that is difficult to model solely by regression analysis into dependent and independent variables.

Like regression analysis, time series analysis is disadvantaged in performance when the relations are not well defined, or data is collected irregularly. This can often be the case for healthcare data, where the relations between the patient, doctors, hospitals and syndromes are not straightforward, but rather complex. The patient does not report to the doctor or become ill at regular intervals, and the available data is not observed in a sequential manner, but rather it is reported when patients report to a service provider. Therefore, the complex and irregular interactions between multiple attributes in healthcare data are often unsuitable for analysis by time series analysis alone.

Network-based prediction models are mostly designed to capture the interrelations of different attributes in complex networks. In the context of healthcare data, they often focus on predicting missing links that were not observed during the transition of the networks. In most cases, these methods focus on a single type of node and a single relation (e.g., authors and their collaboration, or people and their friendship). Compared to the simplicity of this single-attribute network, healthcare networks are vastly complex, having scores of attributes. In addition, traditional network models do not have longitudinal or spatial aspects. Conversely, prediction of disease intuitively depends on the frequency and duration of the patient’s disease history and the corresponding treatment provided, which requires longitudinal timestamps. Similarly, disease prediction depends on spatial aspects of the data. For example, people from different regions and different
socioeconomic backgrounds have different rates of susceptibility to diseases. Also, the spread of disease, especially contagious diseases, is highly dependent on proximity, and follows paths of international and domestic transportation. Traditional network-based approaches often fail to capture these features of healthcare data. Table 2.3 summarises the key properties of different prediction models discussed here.

**Table 2.3: Comparison between different disease prediction models**

<table>
<thead>
<tr>
<th>Criteria</th>
<th>Score-Based Model</th>
<th>Collaborative Filtering</th>
<th>SNA-Based Prediction</th>
</tr>
</thead>
<tbody>
<tr>
<td>Strength</td>
<td>Works well in limited and specific conditions such as ICU or cost grouping</td>
<td>Can predict risk when diseases have strong comorbidity</td>
<td>Can reveal the complex relationships between diseases, patients and physicians</td>
</tr>
<tr>
<td>Dynamic approach</td>
<td>No, benchmarks are defined on strict historical data</td>
<td>Partly</td>
<td>Yes</td>
</tr>
<tr>
<td>Emphasis on chronic disease</td>
<td>No</td>
<td>Partly</td>
<td>Yes</td>
</tr>
<tr>
<td>Utilises electronic and modern healthcare data</td>
<td>Partly</td>
<td>Only few methods have used claim data for prediction (i.e., ICARE system)</td>
<td>Partly: mainly focused on effective collaboration of physician and patients; in some cases worked on the social aspects of disease propagation from healthcare data</td>
</tr>
<tr>
<td>Considers longitudinal nature of disease occurrence</td>
<td>No</td>
<td>No</td>
<td>Partly: the branch of longitudinal SNA considers the temporal component, but has rarely been applied on healthcare data</td>
</tr>
<tr>
<td>Dataset size</td>
<td>Relatively smaller</td>
<td>Large</td>
<td>Large</td>
</tr>
</tbody>
</table>
2.7 Opportunities for SNA and Graph Theory-Based Methods

The current literature on predicting chronic diseases varies widely in terms of implementation and methods. We grouped the existing research into three broad genres and discussed their limitations and fields of strength. This review created opportunities to propose new approaches by extending the concepts of current ones in utilising healthcare data. First, we want to exploit the potential of the electronic healthcare dataset that is gathered by private healthcare organisations. The dataset originates from a group of healthcare insurance providers who came forward as a consortium with the similar goal and unified reporting scheme. This relatively accurate and vast dataset, covering a large proportion of the Australian population, is suitable for analysing relations between diseases and their comorbidity patterns. The dataset is reported electronically and contains quality of service, hospital resource consumption, complexity and comorbidity conditions.

Second, the dataset has timestamp information that can be analysed using the theories of longitudinal social network analysis. This type of analysis can effectively capture the time-situated nature of disease occurrence, which is necessary because the sequence of occurrence of comorbid conditions is important for the prediction of chronic disease. Also, this timestamp data is vital for understanding how diseases occur after certain durations of time, and how these gaps in time affect the overall health trajectory of the patient. This can reveal how, for some patients, taking certain diagnostic tests and preventive measures can lengthen the time after which the same condition will occur again. We can conclude that such preventive measures are vital for promoting remission of chronic diseases over longer periods.

Third, exploring the complex relationship between demography and physicians’ collaboration on the quality of service is another potential for our research. The healthcare system is essentially a complex interaction between different healthcare facilities, physicians and patients with their comorbid conditions. Treatment patterns and collaboration have an effect on patient outcomes and management of chronic disease in
the long run. Traditional methods of analysis often ignore these complex relationships by focusing only on certain aspects of the data, such as comorbidity. Demographic information such as race, sex and geographic location can also have a profound effect on patient outcomes. Incorporating these parameters into our approach can potentially lead to a better prediction method.

A fourth opportunity is to take advantage of advanced coding systems followed throughout the dataset. The earlier methods often had to rely on limited datasets over different formats. The coding system was also not standard, and availability of data from people of diverse geographical locations and socioeconomic conditions created many difficulties, rendering different methods incompatible with one another. At present, to maintain uniformity between various public and private providers, policy makers and government organisations, most governments and healthcare organisations have adopted uniform coding schemes for recording and reporting the available healthcare data. Especially in the last decade, several milestones were achieved in standardising disease codes. For example, ICD has been proposed and adopted by many countries, which has created a common platform for analysing data from a wide range of sources. In addition, with the introduction of database systems, the field of healthcare informatics gained recognition, as analysing large relational networks become possible to do more efficiently. These developments have created opportunities to extend and propose newer methods that can take advantage of these standardised codes and scoring systems.

In this section, we have discussed how improve clinical coding and electronic healthcare data have paved the way to explore the networked property of comorbidity and other healthcare entities. We have mentioned how social network theories (SNA) can potentially be utilised in this context. Now, SNA is a big field and there are different theories that have been proposed and applied to different scientific fields. Before incorporating some of these theories into the context of our framework, we should briefly discuss these theories. Therefore, we do it in the following section.
2.8 Social network theories

Theories in the Social Network Analysis (SNA) are focused on understanding the formation and evolution of social ties (e.g., friendship, collaboration etc.) between actors (e.g., people) in the network. These theories suggest that the influence of entities in a social network is affected by the structural properties such as whom they are connected to, how they communicate etc. Also, these theories try to explain how social networks form, which are the likely entities to form ties in future, which set of entities are forming a social group etc. In this thesis, we focus on the chronic disease progression or network of diseases. Therefore, we need to understand how diseases co-occur (i.e., disease comorbidity) and how their interrelation can be analysed to understand the overall nature of the progression. This can help us to predict the likelihood of a patient’s progression towards chronic disease such as type 2 diabetes. Therefore, we discuss below some of the fundamental SNA theories that will form the basis of the predictive framework.

2.8.1 Effect of network pattern on performance

The performance of the actors within a social network depends on the effective flow of communication (communication pattern) between them. Some communication patterns can be informal (e.g., student network, friendship network) and can emerge, stabilise and evolve organically over time. Conversely, other patterns can be formal or fixed (e.g., organisational chain of command, defence network) where patterns are well defined and imposed. While different kinds of patterns can all yield satisfactory goals (e.g., completion of work, innovation, adaptation), some patterns are more effective than others. How different communication patterns affect overall performance and personal satisfaction (morale) has been a research question since long ago. Harold Leavitt (1949) and Sidney Smith (1950) did some exploratory experiments on this question at MIT, and Bavelas (1950) subsequently reported and analysed their results concerning success rate, network stability, leadership and morale for different communication patterns. The experiment was set up as follows: five persons are placed in five uniquely coloured cubicles (Borgatti, 1997), one in each cubicle. The only way they can communicate with one another is by
passing any number of written messages through a tube that connects two cubicles. The messages are written on stationary of the same colour as the cubicle, so the receiver can differentiate between the senders. The inter-cubicle connections can be changed, therefore creating different possible communication patterns. Four separate patterns are tested: the star (wheel), the Y, the chain (line), and the circle, as shown in Figure 2.4.

![Diagram of communication patterns](image)

**Figure 2.4: Different types of communication pattern**

At the beginning, each person is given five symbols from a set of six, thus keeping at least one symbol in common. The objective is to discover which symbol they all have in common. Each cubicle has six switches on the wall, labelled by each of the possible symbols. By passing messages, the players try to figure out the common symbol and flip the corresponding switch, with the freedom to change decisions until the game is halted. When all five persons flip a switch (i.e., guessing answers), the experiment is halted, answers are checked to decide whether it is right or wrong and the elapsed time is measured. Leadership evolution and morale (job satisfaction) are also identified by asking questions at the end of the experiments. Network stability is measured as the time taken to reach the goal.

The findings from the experiment showed that the success rate was highest in the Y-pattern, then in the line pattern and lowest in the circle pattern. In the original experiment, the star pattern’s success rate was regarded as erroneous due to the confusion of the group members. Results of a survey of participants also strengthened the hypothesis that the probability of leadership emergence is highest at the position of
highest centrality. The middle positions of the line, star and Y structures had the highest occurrences of recognised leaders, whereas the positions of the circle pattern had relatively less deviation of frequency or leadership occurrence. The findings also suggest a direct relationship of morale with the individual’s position within the communication pattern: the more isolated one individual is, the lower their morale. An average morale rating of the Y, star and line patterns given by the most centrally positioned actors was greater than that of peripherally situated actors. Finally, in terms of stabilisation, the star pattern was stabilised earliest, followed by the Y-pattern. The line pattern was less stable than the star, and the Y-pattern and circle pattern were mostly unstable.

In addition to the above findings on the effect of structural pattern on social communication parameters such as leadership emergence, morale, success rate and stability, Bavelas (1950) also hinted that another parameter—innovation—is affected by patterns. Bavelas showed that, as the participants had liberty over the content of the communication, there was an easier and elegant way to solve the problem other than the trivial one. This is to say that the participants could communicate the symbol they did not have rather than communicating the symbols they had. Although following this method was not trivial at first thought, it occurred in all the patterns, and the frequency was fairly even. However, its adaptation occurred mostly in the circle pattern and then in the line pattern. Bavelas concluded that this result may suggest that occurrence of innovation decreases when centrality is highly localised in a communication pattern.

Apart from the human communication network, actor centrality has also been observed to affect performance in different networks, such as road networks or friendship networks (Wang, et al., 2011). Similarly, in a disease network, the disease in the most central position—having connection to many other diseases—can indicate the higher prevalence and significance of the disease. The amount of significance is quantifiable; that is, mathematical formulations of centrality measures exist. Some of these measures are discussed in later sections, but first there is another social network theory that is worth noting in the understanding of disease networks.
2.8.2 Strength of weak ties

One of the earliest studies on how micro-level interaction affects the macro-level structure of a social network was done by Granovetter (1973). He argued that the processes within interpersonal networks translate small-scale interactions into large-scale patterns. He analysed various macro-level phenomena, such as diffusion, social mobility, political organisation and social cohesion, as investigated by different researchers and discussed their results. He emphasised weak ties, and argued that they play a greater role in social phenomena like innovation and diffusion than strong ties.

Many interpersonal ties indeed form large social networks. On the most macro level, if two persons form a strong tie between them, then it is more likely that if a third person forms a tie with either of the two, he will also form a tie with the other one. Further, as time passes, these bonds are likely to become stronger and to form a close triad (Heider, 2013, Newcomb, 1961). Thus people with similar personalities bond together over time, and the strongly-connected triads merge to form a larger group. However, for a social network, this strongly-connected group cannot scale up infinitely, because persons cannot maintain strong ties with an infinite number of people, and also they have certain differences of choice. In this way, a large social network is likely to be formed by different strongly-knitted smaller groups. Now, the groups might not be fully isolated; persons from different groups can also form ties with each other. Granovetter (1973) argued that these inter-group bonds are more likely to be weak ties; this is because, if a strong tie is formed between groups, then the other friends in the groups are also likely to befriend each other, and the partition will eventually collapse. This suggests that if one person wants to communicate with others from the different group, it should be through one of the weak ties. Thus the persons connected by the weak tie act as a local bridge.

Now if a rumour or novel information is to be diffused in the network, then it is through the weak ties by which it spreads further. Using a similar explanation, Granovetter (1973) argued that marginal figures are ‘early adopters’, able to spread the innovation successfully because they might be rich in weak ties. He further explained his weak tie
argument using the results of two experiments. The first one was the ‘small world’ experiment done by Milgram and his associates. In the experiment, Milgram gave packets or booklets to random persons in several United States cities. The task for them was to send the booklet to the person’s address as given in the packet. And, because they did not know the recipients, they needed to send them to some of their friends whom they think could know the recipient or can forward to other correspondences to keep the process running. The experiment was done in several modified ways but the overall principal was the same – to send a booklet to unknown recipient through correspondence and check how many hops it takes to reach. Granovetter pointed out that when the first transfer of the booklet is from a White to a Negro ‘acquaintance’ (considered a weak tie), then the success rate of reaching the destination is 50% (Korte, 1967, Korte and Milgram, 1970). In contrast, that rate falls to 26% when the first transfer is from a White to a Negro ‘friend’ (considered a strong tie). This can indicate that weaker interracial ties are more effective in bridging social distance. The second experiment Granovetter used to explain his results was by Rapoport and Horvath (1961). Some \( N = 851 \) students were asked to write to eight of their best friends in order. Networks were traced by considering two consecutive choices of best friends. It was observed that the most students were found in a network when only the last two choices were considered, and the least were found when the first two choices were considered. Granovetter argued that this indicates that strong ties result in overlapped networks.

Finally, Granovetter (1973) elaborated on the implications of his weak tie argument and the findings of the above described experiments on both micro and macro levels of a social network, citing some empirical results to illustrate these implications. On the micro level, Granovetter’s theory implies that the fewer indirect contacts one has, the more encapsulated he will be in terms of knowledge of the world beyond his own friendship circle. Granovetter developed this point empirically by citing some results from a labour market study where he investigated how workers find new jobs through personal contacts as opposed to any other methods. He found that, in most cases, the workers (55.6%) found their jobs through contacts whom they saw occasionally. Also, the path length of
the information flow (i.e., number of intermediaries) in most cases (45.3%) was 1, thus strengthening his weak tie claim.

On the macro level, Granovetter (1973) discussed why some communities organise for common goals effectively, whereas others seem unable to adapt to doing so. As an example, he cited the Italian community of Boston’s West End, who were unable to fight against urban renewal and were ultimately destroyed. He argued that the West End community might have been partitioned into isolated cliques, resulting in a lack of significant communication paths due to the absence of weak ties between the groups. Thus trust relationships between leaders and citizens were hampered, rendering unsuccessful the campaign against the urban renewal.

2.8.3 Freeman’s concept of centrality

Researchers have long been trying to discover how to quantify the level of performance of actors in a social network in terms of characteristics like innovation, efficiency and leadership. The most intuitive approach to this investigation has been via the concept of centrality, which captures the sense that some actors of the network are more central than others, and hence they differ in measures of performance. Although the precise effect of centrality on one’s performance and how to quantify that centrality measure is still an open research question, the idea of it was first introduced as early as the 1950s. A number of experiments done on centrality (Bavelas, 1950, Burgess, 1969, Cohn and Marriott, 1958, Czepiel, 1974, Pitts, 1965) suggested that centrality is relevant to group performance. However, results were confusing and contradictory (Burgess, 1969) in terms of the definition and quantification of centrality.

Freeman (1979) wrote a seminal paper clarifying the concept of centrality in social networks and its applicability and limitation in group performance. He reviewed old centrality measures, fixed some gaps and expressed the measures in terms of a common set of symbols and graph theory. He divided the measures into two contexts: one from the perspective of a particular actor of the network, known as point centrality, and the other
from the whole network level, or graph centrality. Point centrality deals with the structural importance of a particular actor within the network, while graph centrality expresses the overall compactness of the whole network or graph.

Regarding point centrality, the measure of importance of a particular node depends on the context of the experiment. The measure can be how much communication the actor is doing with neighbours, how close it is to others or how strategically it is positioned among others. Based on these criteria, three centrality measures have been proposed for actors: degree, closeness and betweenness. Degree centrality is measured by how many direct neighbours an actor has, closeness is defined by how closely it lies to all other actors in the network, and betweenness is measured by the extent to which an actor lies on the shortest path to all others in the network. These centrality measures indicate the varying structural importance of an actor within the network: degree centrality represents the activeness of an actor, betweenness indicates the potential of control and closeness specifies independence or efficiency.

Freeman proposed two versions of each of these three point centrality measures. The first version is dependent on the network size, while the second version is normalised in the range of 0 to 1 so that actors of different network sizes can be measured and compared on a uniform scale.

Freeman also measured similar forms of centrality on the network level. The network-level measures are based on aggregations of the same measures on the actor level, and finding averages. For inter-network comparison, there is a normalised version of each of the three measures.

Freeman finally applied his centrality measures on all possible graphs comprised of five points and compared them. He showed that all three measures give maximum scores to star or wheel networks, and minimum scores to circle and complete graph networks. Measures on all other possible networks lie somewhere between these archetypes, except for disconnected networks, which cannot be measured. However, these
intermediate measures are distorted in their relative ranking. The distortion gives rise to another question on how these variations of centrality interpret when an actor changes its structural position within the network.

### 2.8.4 Structural holes theory

Burt (1992) introduced the concept of a structural hole in explanation for how, in a competitive social network, an individual’s performance and entrepreneurship opportunity greatly depends on access to non-redundant information. He emphasised structural positions rather than network structure and relational ties. Earlier studies often assumed that the strength of interpersonal ties does not change over time. For example, Coleman et al. (1966) stated that strong ties in a full closure network, i.e., strongly connected network, foster social control of deviant behaviour. Here it is assumed that each tie is a provider of unique information or communication. However, these assumptions often led to contradictory empirical results that in some extremely dense networks with strong ties, individuals failed to perform better.

Burt (1992) gave an explanation as to why some individuals can perform better than others with a network model of competition. He further elaborated Coleman et al.’s (1966) study and assumptions on information diffusion within a network from a structural point of view, rather than one of interpersonal relations. The key element of his theory is structural holes, which are the mechanism underlying Granovetter’s (1973) claim that weak ties are more useful because they give actors access to novel information. As discussed previously, a large social network is composed of several closely knitted groups. Between the structural holes of the groups, actors who work as brokers for passing information are structurally in a better position in terms of access to novel information.

Burt argued that there are several reasons that made the brokerage position structurally advantageous, as opposed to a position within a closely knitted network. First, in a social network a person cannot maintain interpersonal ties consistently over an extended period of time. Keeping a tie incurs some maintenance cost, and thus a large number of
interpersonal ties are not feasible for an individual (Gurevitch, 1961). Second, as a dense network grows, the information becomes more redundant. The number of retellings of the same information creates a damping effect. Therefore, in practice, a social network is structurally divided into several clusters or cliques. Burt introduced his ‘structural hole’ concept to describe the absence of strong ties or gaps between these groups. These holes create social capital via brokerage opportunities. He argued that the actor who sits in between these clusters, thus bridging up the gaps between the structural holes, gains competitive advantages in several ways in terms of brokerage position, economic gain or controlling capability.

2.8.5 Homophily and cluster

Most real-world networks do not generate or evolve in a random manner. Given a set of students previously known to each other, who join a new course, they will eventually form a friendship network among themselves. The ‘who chooses whom’ preferences will not arise randomly. Several factors can affect the choices: for example, friendliness or academic performance can have an impact on the friendship. The tendency to form interactions based on similarity in a social network is called homophily (McPherson, et al., 2001). In disease networks, as we discussed earlier (see Section 2.4, ‘Networked Approach in Healthcare’), different diseases may share common biomarkers or environmental factors. Those diseases are likely to occur together: that is, they follow the homophily concept. For example, hypertension and T2D are related: insulin resistance and diabetes can precipitate hypertension by stimulating the sympathetic nervous system and the renin–angiotensin system, and promoting sodium retention (Lago, et al., 2007). Therefore, if a person has hypertension, they are at risk of being diagnosed with T2D. Thus, we observe preferential attachments between related diseases in the disease network.

Now, if the network follows preferential attachment rather than random formation, it may form a cluster. A cluster within a social network is a group of similar actors that have a great deal of interaction with each other and fewer interactions with the actors outside of the cluster. In this thesis, we form a disease progression network where we also look at
the clusters of diseases that frequently occur together. We then try to investigate whether a patient has diseases that overlap with the clusters. Different mathematical functions and algorithms exist for detecting clusters or communities in a social network; we use the one proposed by Blondel et al. (2008), as it is suitable and reasonably accurate for large networks. The detailed methodology for that is discussed in the later chapter.

2.9 Summary

In this chapter, we have reviewed the existing literature and set up a clear background on which the framework is introduced. We have discussed related theories and methods that are being used in understanding chronic disease progression and prediction of disease risk. We began with a brief introduction to prediction methods and evolution in healthcare and discussed how computer technology has disseminated in this field, giving rise to the concept of electronic healthcare or e-health. We then discussed different groups of methods and their implementations in disease prediction. This was followed by a discussion of the limitations and opportunities found in the existing literature. Finally, we presented some fundamental social network theories that will form the basis of our disease prediction framework.

In the next chapter, we introduce the context of the research. As we implement our framework on a real-world health dataset, it is essential to understand the data flow process and the different components of the data that are used during the implementation and analysis. These can vary considerably across geographical locations employing different legislation, policy and technology. Therefore, before going over the methodology in detail, it is essential to properly define the context from which the data comes. For our research, this is the Australian health system, specifically administrative data received from private health funds. The context-specific features, as well as the basic structure of the framework, are discussed in the next chapter.
Chapter 3: Contextual Background of the Administrative Dataset and the Framework

In this chapter, we introduce some context of the dataset and the framework in a comprehensive manner. We begin with a brief introduction to the Australian health system, followed by a detailed description of the data collection process and data propagation within the Australian health system. Next, we discuss different practical issues with the healthcare dataset that we need to take care of during the data curation and filtering process. Following this, we addressed the data security and privacy issues. Next, we discuss the organisation of the dataset. Then we introduce some core concepts that form the basis of the framework. Finally, we give an overview of the basic structure of our predictive framework.

3.1 Australian Health System

The Australian health system is complex and multifaceted. It can be described roughly as a complex web of components interconnected by multiple service relations. A broad-level categorisation of these healthcare components would give us three major entities, namely (a) consumers, (b) providers and (c) policy makers and funders.

Consumers are the recipients of healthcare services and remain the focal point of the overall system. The term ‘consumer’ is not restricted to Australian citizens, but rather encompasses overseas visitors, temporary and permanent visa holders and asylum seekers (AIHW, 2014). Consumers finance their healthcare services through a combination of three major types of funding, namely (a) Medicare, (b) private health cover and (c) out-of-pocket pay. Medicare is the universal government-funded healthcare scheme provided to all Australian residents and certain categories of visitors. All eligible consumers receive free or subsidised (as per the Medicare Benefits Schedule or MBS) treatments under Medicare cover. Medicare Australia also administers the Pharmaceutical Benefits Scheme (PBS), which partially covers the cost of certain prescription medicines. Consumers can opt
from a range of private health insurance funds to cover the cost of services that are not covered or partially covered by public funding. Private health cover owners also have greater freedom to choose doctors and schedule non-emergency surgeries. Often, public and private funding cannot cover the full cost of the service. In these cases, the remaining ‘gap’ is paid by the consumers as an out-of-pocket fee. According to the AIHW, in 2013–14 total health expenditure in Australia was $154.6 billion, which translates to $6,639 per person (AIHW, 2015). Figure 3.1 shows the different funding sources that met this expenditure.

![Figure 3.1: Different funding sources for total health expenditure during 2013–14](image)

Providers are the trained workforce or service settings responsible for delivering healthcare to consumers. The category includes medical practitioners, nurses, allied and other health professionals, hospitals, clinics and government and non-government agencies (AIHW, 2014). Providers can work in the public or the private sector. Policy makers and funders are the people or institutions in charge of administration, healthcare policy making and overall funding. They include but are not limited to the Commonwealth, local, state and territorial governments and health authorities, and private health insurers.
Consumers, providers, policy makers and funders are strongly embedded in the health system through a plethora of relations. Patients may choose to visit different specialists, general practitioners (GPs), hospitals or other providers during their lifetime for different health conditions. Providers within a region need to communicate together to deliver better service to the community. Policy makers and funders need to monitor the overall performance of the system in terms of quality of care, value for money and budget. They also track the health status of the community as well as the whole population on a broad level. The overall picture of the healthcare entities and their interrelations is complex. We present a simplified version of the health system in Figure 3.2.

![Figure 3.2: Organisation of Australian health system](image-url)
3.2 Hospital Casemix Protocol Dataflow and Structure

The Hospital Casemix Protocol (HCP) is the data specification that is maintained throughout the pipeline of data transmission, starting with data generation and ending as a research dataset, through different stakeholders and agencies. At its origin, providers regularly collect data from patients when they access healthcare services. These data are used for record keeping and performance monitoring. They may also be shared among other providers in the network or reported to stakeholders for auditing, management or research. Types of healthcare data vary widely depending on the kinds of service, providers and consumers involved. For our research context, we focus on administrative data from hospital admissions. This type of data is generated to classify admitted patients into different groups based on their diagnoses, associated procedures, complications and other healthcare matrices (e.g., length of stay, overall cost). Patients of the same group have similar footprints in terms of health status and resources required; in other words, they form clinically homogeneous groups that are expected to consume similar amounts of resources. This process is described by three related terms:

1. The concept of grouping the patients into clinically homogenous groups is termed as *Casemix* (Department of Health, 2013).

2. The data specifications in which the Casemix should be recorded and transmitted to stakeholders is called the *Hospital Casemix Protocol* (Department of Health, 2016), commonly known as HCP. The dataset that we use in this research essentially originates from HCP data.

3. After applying Casemix classification on the patient record, the patient is grouped into a single *AR-DRG* (ACCD, 2016) code, which stands for *Australian Refined Diagnosis-Related Group*. Details of AR-DRG are discussed later in this chapter.

In our research context, the HCP dataset includes clinical, demographic and financial information for patients who are members of private health funds. HCP data collection for privately insured patients officially began in 1996–97 (Department of Health, 2016), and
the specification updates periodically. The data originates in hospitals and is transmitted to respective private health insurance funds, who then send the data to the Department of Health. Therefore, HCP has two major specifications: one for the hospitals to submit to the insurers, and the other for insurers when they send the data to the government (i.e., the Department of Health). The specifications are extensive and take into consideration all possible services and circumstances that may occur between admission and separation. The full HCP specification goes beyond the scope of our research, and therefore we only introduce the data items that are relevant later on in the Section 3.5. Up-to-date HCP specifications are usually available from the Department of Health website.

3.3 Factors Affecting Coding Quality in the Dataset

Healthcare datasets are collected over a span of several years in different healthcare facilities. In most cases, the data undergoes several layers of processing, including recording, formatting, linking and transferring between various agencies before it finally reaches the researchers. Moreover, during the data collection period, the policy and specification related to the overall handling process of the data may change multiple times. This may introduce heterogeneity into the data, and can introduce different types of errors and inconsistencies. In formulating the research method, we should adequately consider these issues and ensure consistency in the data for the best possible outcome. We briefly discuss the major issues that contribute to the overall coding quality of a dataset in the following subsections.

3.3.1 Expertise of the coder

In most healthcare settings, trained healthcare professionals collect the administrative data. These professionals are often designated as ‘clinical coders’, whose job is to translate information from a patient’s medical record, which in most cases consists of written clinical documentation, into an agreed healthcare data classification system (WA Health, 2016). In Australia, the ICD-10-AM/ACHI/ACS coding scheme is used to classify healthcare data (AIHW, 2016). The three components of the coding scheme are
abbreviated as follows: the International Classification of Diseases 10th Revision Australian Modification (ICD-10-AM), the Australian Classification of Health Interventions (ACHI) and the Australian Coding Standards (ACS). A clinical coder can also be assisted by a software tool (e.g., DRG Grouper) to record or transmit the data.

The overall quality of coding depends on several factors related to the coder. The coder’s expertise in the form of training and experience, as well as work environment within the healthcare facility, may affect the coding quality. For example, ICD codes vary in length: adding a suffix to the core part of the code can provide further information about the diseases, such as their aetiology, anatomical site and severity (Centers for Medicare and Medicaid Services, 2016). If a person has type-2 diabetes with ketoacidosis, the coder should record E11.1 for that condition. However, if the coder simply records E11, which only indicates type 2 diabetes, they are providing impoverished information: there will be no way for researchers to tell which specific type of diabetes mellitus the patient had. To give another example, assume that the patient has gestational diabetes mellitus and the onset arose during pregnancy, not pre-existing. The ICD core code for this condition would be O24.4. There may be a chance of miscoding this condition as E11.x, the generic code for type-2 diabetes, in which case it would not be a correct coding. Such coding discrepancies can also occur due to lack of availability or access to the reference manual, which may be a medical dictionary, codebooks or software. Further, several other factors, such as the version of any coding software used, motivation or time constraints, can also play a role in the level of coding.

### 3.3.2 Policy

Healthcare rules and regulations encounter frequent updates. As healthcare data is collected over several years, any midterm policy changes can incur visible pattern changes in the trends and may cause significant inconsistencies within the dataset.

As a real-world example, we can mention the policy changes for the coding practice of type-2 diabetes in New South Wales (NSW), Australia (NSW Health, 2016). In July 2008,
the Australian health policy for hospital data concerning the coding of additional diagnoses for diabetes was revised. The revision was made in response to the increased relevance of diabetes to the treatment for the principal diagnosis and the length of the hospital stay. This resulted in a requirement that coders include diabetes as a comorbidity only if a substantial alteration to the clinical treatment regime took place during the hospitalisation due to diabetes. This policy change had a significant impact on the numbers and rates of diabetes as a comorbidity only (not as a principal diagnosis). For example, in the NSW hospital statistics, a 54% reduction in rates of disease involving T2D as comorbidity is observed between 2007–08 and 2008–09.

The Australian Coding Standard for diabetes was revised again in July 2010. This resulted in a major change, mainly affecting the coding of diabetes as a principal diagnosis and, to a lesser degree, the coding of diabetes as an additional diagnosis or comorbidity. According to the new policy, the coding of the principal diagnosis was to take place first (without reference to the fact that a patient may have diabetes), and additional diagnoses were then to be coded according to the guideline of relevance and the impact of diabetes on the clinical management and treatment of the principal diagnosis. For example, an admission for the treatment of a cataract or ulcer in a patient with diabetes before 2009–10 would have been coded as diabetes with the specific complication of an cataract or ulcer (NSW Health, 2016). However, after July 2010, a cataract or ulcer was coded as the principal diagnosis, with diabetes only appearing as an additional diagnosis if it made an impact on the hospital’s management or treatment of the cataract or ulcer. This particular change caused a dramatic 60% drop in the number and rate of hospitalisation for diabetes as a principal diagnosis in NSW between 2009–10 and 2010–11. It also contributed to the existing trend of decreasing comorbidity rates incurred from the previous policy change.

A further health policy change took place in July 2012. According to this new rule, all health conditions that a patient has at the time of hospitalisation should be recorded as additional diagnoses, regardless of their impact on the main cause of hospitalisation. This
change partially contributed to a rise of 4.7 times (or by 370%) in terms of admission with diabetes as comorbidity, between 2011–12 and 2012–13; this trend has since stabilised.

It is evident from the above examples that policy changes can introduce consistency issues within a dataset. While formulating the research method and framework, we should be careful about taking the policy change issue into consideration.

### 3.3.3 Funding

Funding may also affect the quality of coding. Better funding structure for the medical documentation, reporting and coding system will translate into better training, policy implementation and facilities like software and IT equipment. This, in turn, will result in better quality of coding.

### 3.3.4 Coding standards

Coding schemes like the Commonwealth Medical Benefits Scheme (CMBS), ICD, and DRG change over time. Older coding standards often lack precision or may not precisely specify a disease. Coding standards are also frequently updated to keep up with recent advancements in diagnosis or pricing areas. However, the implementation of newer coding standards requires sufficient policy revisions, funding, training and updating of software and equipment, which takes time. As a result, several coding versions may be present in the same dataset concurrently over the years. Such inconsistency needs to be taken care of during the data preparation process.

### 3.4 Healthcare Data Security and Implementation

Healthcare data contains personal information and medical history of patients. This makes it highly sensitive and raises the issues of privacy and confidentiality. Therefore, governing bodies in charge of collecting, organising and transmitting the data impose different levels of cautionary and security measures to ensure that the privacy is protected under all circumstances. While the full spectrum of privacy and security measure is extensive, we
briefly identify the major issues among them which are particularly related to our research context. We discussed them in the following sections.

3.4.1 De-identification

Data fields like patient name, postcode, contact number, address and identifiers (e.g., Medicare, social security number) can potentially identify a patient when they are present in the dataset, either individually or in combination. While these fields are often necessary for service providers, governments and insurance companies, for research purposes they are not usually required. Therefore, personal information is often deleted before transmitting the data to another party (i.e., researchers or data scientists). In some cases, de-identification is done partially to keep some information intact while minimising the risk of security breach: for example, the exact date of birth is omitted and the only birth year is mentioned. Similarly, sometimes adjacent postcodes are merged to identify greater geographical areas to minimise the risk of identification. In other cases, data is aggregated.

De-identification can have limitations. If the researchers have another dataset for the same or somewhat overlapping cohorts, they may need to link the two datasets to form a complete cohort. For example, if the researchers have GP data and receive hospital admission data for the same cohort, they may want to match the records to see the primary and tertiary care matrices for the patients. This would not be possible if the received data were de-identified in a manner that is not re-identifiable. To mitigate the problem, the original provider can replace the personal identifiers with randomly generated unique IDs and send the de-identified data to researchers. At the same time, the provider can keep a translation table of personal identifiers against the unique IDs that were given to researchers, so that, if needed, the provider can re-link the result from the researcher’s end with their original data. Also, if the policy permits, the translation table may be used to link up with other data sources that have the same or overlapping cohorts. The following Figure 3.3 shows an overview of the data de-identification process.
3.4.2 Scrambling and encryption

Data scrambling refers to the process of obfuscating or removing sensitive or personal data (ORACLE, 2014). This is a part of the de-identification process, and is irreversible so that the original data cannot be derived from the scrambled version. Encrypted data, in contrast, is reversible in the sense that with the proper decryption key, the original data can be retrieved. Encryption is usually done in order to protect the data from unauthorised access, so that in the event of a breach, the actual data cannot be derived by unsolicited users. Only intended users will have access to the decryption key or software that can retrieve the actual research data.
3.4.3 Data access control

Healthcare data that is not an aggregated derivation of individual patient records is not normally made accessible to the general public over the internet. Even if the patient records are de-identified, there is still a chance of re-identification. For example, if there is a very small population of a certain age group or having particular disease within a geolocation, it might be possible to identify a few patients based on their age, postcode or disease code if present in the dataset, without the help of specific personal identifiers. For these reasons, healthcare data is generally hosted in a secure software and hardware environment. Hardware implementation of data access control includes a dedicated, secure server system inside a closed research facility. Software implementation may include firewall, VPN (for remote access), encryption and antivirus programs, and most importantly, password protection allowing only specific users to access the system. Further, access may be restricted to dedicated workstations to reduce the risk of security compromise by unsolicited users or software.

3.5 Dataset Entities

As discussed, healthcare data can be stored, represented or transmitted in different formats in terms of database structure, naming, level of detail and security. One of the main focuses of this thesis is to present its framework a uniform way so that it can be transformed and implemented in different healthcare settings and on different datasets. Nevertheless, the underlying organisation of the dataset should follow standard relational database design best practices for efficiency and consistency during analysis. Therefore, regardless of the initial organisation of the dataset, we should reorganise the dataset into a suitable relational database format. The primary entities of the dataset as used in our present analysis are described in following subsections. The database organisation is shown in the form of an Entity-Relationship (E-R) diagram in Figure 3.4. A more technically detailed version of the E-R diagram is included in the Appendix E for reference.
3.5.1 Patient

In the originally received dataset, this entity was named ‘Members’. The reasoning behind this naming is that the data was received from the private health funds, and from their perspective, consumers need to purchase membership to have access to private health insurance. Different types of membership exist; for example, one can buy a single cover for oneself, or a family cover for all family members. The cover can also be ‘basic’ or contain insurance for ancillary or other services. To reflect these variations in membership type, the original dataset had two records: ‘membership_ID’ and ‘member_ID’. Everyone under the same family cover had the same ‘membership_ID’, but each had a different ‘member_ID’. Therefore, to point uniquely to a patient, we needed to combine these two IDs together.

In the data preparation phase of the analysis, we simplified the member and membership IDs into one unique ID as a primary key, as we do not need to know whether an individual
has family cover or single cover. This ID was named ‘pntID’. Other mentionable fields storing patient information in the table include year of birth (‘pntYoB’), membership date (‘pntMembershipDate’) and sex (‘pntIsMale’: if the value of ‘pntIsMale’ is 1 the sex is male and otherwise female). The dataset does not contain any personally identifiable information such as name. Two other fields that are used to store calculated information later on are ‘pntCohort’ and ‘pntFirstDiagDate’. The ‘pntCohort’ field contains the information about which cohort the patient belongs to; the use of cohort number is discussed in the ‘Methods’ chapter. The ‘pntFirstDiagDate’ field contains information on the earliest date when the patient is diagnosed with the particular chronic disease, if any. The value is kept null if the patient is non-chronic.

3.5.2 Providers

This entity contains the information about the healthcare service providers. A provider can either be a health professional (e.g., specialist doctor, nurse or anaesthetist) or entity (e.g., hospital). Most traditional service providers are identified by a unique code provided by the Australian Medical Association (AMA), but for non-standard health services (e.g., acupuncture or homoeopathic medicine) the ID may not be standard. In our analysis, we do not require any particular provider to be tracked or linked by their AMA code. Therefore, we only made sure that the provider ID (‘pvdID’) is unique. Also in the originally received dataset, there was a field for each provider named as ‘provider group’. In Australian healthcare, some hospital or clinical facilities belong to large conglomerates, so providers were identified as belonging to a hospital group. We removed these fields from our dataset. Another field is connected to the provider entity is provider type (‘pvdType’). This field contains categorical values that indicate whether the provider is a physician or a hospital: For example, if the field has value 1, it indicates that the service is provided by a physician; if the field has value 2, it indicates that the service is provided in a hospital, and so on.
3.5.3 Admission

Admission is an important entity for our framework. Before describing this entity, we should delineate between the related terms of ‘admission’, ‘separation’ and ‘episode of care’, as in the Australian healthcare context their definitions are slightly complicated based on whether they are used for formal or statistical purposes. An admission is the event in which a patient is admitted to the hospital. This can happen in several ways: the patient can be admitted for a pre-planned treatment or procedure; the admission can be unplanned (e.g., visit to the ER department); or the patient can be transferred from another healthcare facility. In contrast, a separation occurs when the patient is discharged to go home, is transferred to another healthcare facility (e.g., hospital, rehabilitation, nursing home) or dies. Now, for statistical purposes (e.g., administration, audit, billing), the full duration of continuous stay in the same facility can be divided into several admissions and separations in turns based on the type of care provided. In other words, if a patient’s care type changes during a hospital stay (e.g., shifted from the ER department to general departments for secondary treatments), it is preferable to keep the duration of each care type statistically separated. The summary information generated within this period is saved in the admission entity. Therefore, each time the type of care changes, a separation and then another admission event occurs. The duration between admission and separation events is called an episode of care, which is formally defined as the period between admission and separation that a person spends in one hospital including leave periods not exceeding seven days. Each episode is recorded in the dataset as a row in the admission entity. Because in most cases, there is no important difference between the terms ‘admission’ and ‘episode of care’, for simplicity, we will use the term ‘admission’ to refer to an episode of care.

An admission entity has a unique ID (‘admID’) to identify that admission. It also has a foreign key (‘admPntID’) pointing to the patient ID field in the patient table. This keeps record of which patient an admission belongs to. To record the time-related information, the admission entity has two fields denoting time of admission (‘admDoAdm’) and time of
discharge (‘admDoSep’); the gap between these two times is saved into another field called length of stay (‘admLoS’), expressed in whole days. Each admission is associated with exactly one DRG code (see Section 3.5.4). For design purposes, the dictionary of DRG codes is saved as another entity and a field in the admission entity (‘admDRG’) acts as a foreign key pointing to the associated DRG code in the DRG entity.

### 3.5.4 Diagnosis-Related Group (DRG)

The DRG is a coding system designed to classify hospital cases into different groups that are expected to use similar amounts of hospital resources in terms of length of stay, cost, treatment complexity and care. Admitted patients with the same DRG grouping are expected to be clinically homogenous; that is, they exhibit similar resource usage and incur similar costs on the health system. For example, elderly osteoarthritis patients admitted for knee replacement are expected to incur similar footprints on hospital resources and, therefore, they may fall under the same DRG group. The grouping is not, however, done on the basis of mere speculation or assumptions. There is a ‘grouper software’ that takes into consideration different parameters and classifies the DRG based on predefined rules. The parameters include the patient’s demographics (e.g., age, sex), diagnoses (e.g., ICD codes) and procedure codes present in the admission. The whole process of supervised by a trained clinical coder.

The DRG version used in our dataset is AR-DRG. Only one DRG is associated with each of the admissions. To ensure performance and conformity to database design best practice, we put all the unique DRGs into a separate DRG table, thereby creating a dictionary of DRG codes. A foreign key from the admission table then points to the DRG record in the DRG table that is associated with the admission. The table has a unique primary key called ‘drgID’ and the main text (‘drgText’), denoting the actual DRG code. It also has a computed field called ‘drgFreq’; this field essentially keeps track of the number of admissions that have this particular DRG code.
3.5.5 Treatment codes

Depending on the nature of the admission, a patient can undergo different types of treatments. For example, if the admission is due to a simple condition, the record may only contain diagnosis codes in the form of ICD-10 codes. If the patient needs procedures to be performed (e.g., knee surgery), then the performed procedure will be recorded using its own ACHI code. Further, all related fees are charged against different CMBS item codes; the Commonwealth government subsequently schedules payment and benefit against each CMBS items code. All of these different types of codes may be present in the patient’s HCP data, and these are represented in the Treatment codes table. It has two important fields: the first, named ‘treatCode’, contains the exact code. The second field, ‘treatCodeType’, indicates the type of code (e.g., ACHI, CMBS or ICD) in terms of categorical values.

Another relational table, named ‘relAdmTreatPvd’, is used to record which treatment codes in the HCP data are associated with which admissions, as well as the provider of the service. This table has three relational columns. The first one (‘relAdm’) points to the associated admission record in the Admission table. The second field (‘relTreat’) points to the associated treatment code record in the Treatment table. Finally, the third field (‘relPvd’) refers to the appropriate provider’s ID in the Provider table.

3.5.6 Additional derived entities

Apart from the primary dataset entities discussed above, we also have two important entities required for the networked representation of the disease progression. These two entities are derived from the main database through calculation. We now briefly discuss these two entities. Also, the following Figure 3.5 shows the graphical representation of the additional entities.
3.5.6.1 Node

This represents an additional node entity that is used to save comorbidity information about the baseline networks (discussed in detail later in this chapter, section 3.6.4). Node entity has no direct linkage in the forms of ‘primary key’–‘foreign key’ with the main part of the dataset, as this is primarily used for analysis as well as storing the output in the later stage of analysis in the framework. The entity has three major fields: id, label and weight. The ‘id’ field is the primary key of the table; ‘label’ contains the name of the disease or disease group for analysis; and ‘weight’ indicates the prevalence of the disease or disease group.

3.5.6.2 Edge

Like the node entity above, this is also an additional entity. It supplements the node entity, and together they store the information regarding the comorbidity or baseline network (discussed in detail later in this chapter, section 3.6.4). As the name suggests, the edge entity contains the link information about the baseline network. Each row contains information on a single link. There are five fields: id, source, target, frequency (‘freq’), time gap (‘time_gap’) and standard deviation (‘stnDv’). Source and target fields are foreign
keys that point to the ‘id’ field of the node entity which represents the nodes that are connected by this particular edge. ‘id’ is the primary key of the edge entity. The rest of the fields hold the information about the edge attributes of the baseline network.

3.6 Core Concepts of The Framework

The goal of this thesis is to develop a framework for predicting chronic disease risk using administrative data that can perform better than the existing methods. In developing this framework, we incorporate methods from graph theory and social network methods. From the discussion of the literature above, we can observe that a broad spectrum of network-based methods is at work in predicting disease risk or understanding disease associations with underlying biomolecular factors. However, we also mentioned that these existing methods often have limited precision or application over the type of administrative data that we are working with. Therefore, to formulate a framework that can overcome these limitations, we must first examine the features of the dataset being used, and understand the potentials as well as the pitfalls.

Administrative datasets are usually maintained by providers or stakeholders such as government and health insurance companies. The predictive framework, therefore, must focus on the benefit that these stakeholders stand to gain from it. One potential contribution is that identifying high-risk patients using the framework could help the stakeholders tremendously in formulating policy and encouraging those patients to take preventive measures, thus saving cost in the long run. The framework can indeed provide better and more efficient solutions compared to the ways in which those goals are currently approached.

However, while formulating the framework, we should be cautious not to be overconfident in reporting the output. Chronic diseases are extremely complex, and no framework can predict perfectly whether an individual will develop the disease in future or not. So, rather than aiming for a prediction model that will give a categorical ‘yes/no’ answer for the question of future diagnosis, we focus on finding the probability or risk of
chronic disease for a group of non-chronic patients. In this way, the framework can make a more realistic assumption and prediction performance on risk. The stakeholders, then, in turn, can use it to rank their patients in order from high risk to low risk.

In the following sections, we discuss some discrete features that will make up different components of the framework.

3.6.1 Determining research scope: chronic diseases and comorbidity

Administrative data contains information about diagnosis and treatment of patients during hospital admission. Our focus in this thesis is to understand disease progression and develop a predictive framework for that disease. As a first step, we pose the questions: which set of diseases should we focus on? First, the intuitive direction should be to focus on the diseases whose risks can be predicted. Some diseases or diagnoses can occur by mere chance. For example, physical injuries can occur at random, while the pathogenesis of many idiopathic diseases is not understood properly; therefore, we should not focus on these conditions. Besides, we are not interested in prediction based on a spatial analysis; therefore, communicable diseases are not of interest. Second, as stakeholders provide the administrative data, we should be focused on diseases that have central importance in their policy, and the outcome (e.g., the framework) of the research should be translated into knowledge that can help ensure better policy and service by the providers. Third, the focus should be on diseases that are more prevalent within the healthcare system and impede the quality of life of patients. Finally, we should focus on a disease that is properly recorded in the dataset, often a concern for administrative data. Considering these directions, we choose to explore chronic diseases, and more particularly type 2 diabetes (T2D).

Chronic diseases (e.g., diabetes, COPD, cancer) are the group of diseases that tend to be long-lasting and have persistent effects (AIHW, 2016), either over the lifetime or through recurrent episodes of relapse and remission (e.g., Crohn’s disease). In the Australian context, chronic diseases exert a significant social, economic and health burden, and thus
demand a great deal of research attention. In planning research on chronic diseases, we need to determine whether it is indeed worthwhile to predict risk for these diseases. In favour, we bring up the topic of comorbidity, which refers to any two or more diseases that occur in one person at the same time (AIHW, 2016). Research literature and empirical statistics have shown that chronic diseases have a comorbid nature (Schellevis, et al., 1993). This is because these diseases either share similar pathobiology, or are connected to similar environmental or behavioural factors (e.g., smoking), or both. For this reason, chronic diseases are often analysed using the networked approach discussed in the earlier chapter (Section 2.4). Administrative data records the comorbid conditions of a patient (if any) during admission. For predictive modelling, these characteristics make chronic diseases a suitable focus of our research.

3.6.2 Elixhauser index as comorbidity measure

As discussed in the previous chapter (Section 2.3.1: Score-based models), diagnoses and comorbidities are recorded using ICD codes. The version of ICD used may vary depending on country and policy codes; in the Australian healthcare context, administrative data is recorded using the ICD-10-AM version. One of the challenges associated with this method of coding is that many ICD codes present in the data are unrelated to the researcher’s task. For example, the ICD code Y92.01 indicates that the medical condition occurred in outdoor areas; this information is not relevant to include in our analysis. Similarly, there are many ICD codes indicating conditions that are not useful for our task of prediction (e.g., W10.9: Fall on and from other and unspecified stairs and steps). As a result, in developing a framework, we should exclude these unrelated ICD codes, which would create noise within the comorbidity network and affect its precision.

To filter out irrelevant codes, we examined the literature for comorbidity measures (see Section 2.3.1 for details) because they list the important associated conditions that are present during hospital admission. Not all comorbidity measures factor in the same number of comorbidities. Some are more extensive than others; in addition, the measures differ on the quality and number of hospital record sets that were empirically analysed to
develop them. After studying these factors, we chose the comorbidities defined by the Elixhauser index, as it is better empirically validated, shows better performance in risk-based scoring (Sharabiani, et al., 2012) and does not automatically aggregate individual comorbidities, as other measures do. The Elixhauser index comorbidities were originally designed to assess risk of hospitalisation, not exactly to predict the risk of a particular chronic condition. Therefore, using the index as a basis, we further reviewed the literature to find other significant comorbidities and added to the list as necessary. A detailed description of the comorbidities used is introduced in the Chapter 4:

3.6.3 Appropriate coding criteria

The fundamental unit of data for this research is the disease code. In our chosen administrative dataset, these codes are recorded at the time of hospital admission, when physicians diagnose a health condition. Disease codes represented in the dataset can vary widely for various reasons. The data is normally collected from different healthcare facilities over a span of several years. In most cases, the data undergoes several layers of processing, including recording, formatting, linking and transferring between various agencies, before it finally reaches the researchers. Also, during the data collection period, policies and specifications related to the overall handling process may change multiple times. This may result in heterogeneity in the data, and can introduce different types of errors and inconsistencies. For the most part, the disease codes in our dataset are recorded in ICD format. However, there are different versions of ICD coding standards based on the time recorded and the country-specific adaptation of the coding format, and national or local policies may also include the presence of other supplementary disease code formats in the dataset. To derive the framework, we must identify an appropriate coding scheme and any inconsistencies or errors within the dataset. Sometimes it may be necessary to translate the codes to ensure uniformity. The overall choice of coding scheme will largely depend on context and implementation. However, regardless of context, any disease code-related filtering should be included in the framework during the data preparation step.
3.6.4 Representing the health trajectory of particular chronic disease patients

We have already discussed the networked nature of administrative healthcare data. An individual patient can be admitted to hospitals several times during his or her lifetime. Each time the main reason for the admission is documented as the principal diagnosis. At the same time, the patient can also have other complications that may act as underlying causes of the principal diagnosis. Similarly, the patient may have pre-existing medical conditions that are worth noting down for both coding and treatment purposes. To give an example, if a person is hospitalised for COPD, the clinician may write J44 as the principal diagnosis, because COPD is the main reason for hospitalisation. Now, the patient may also have asthma, a related disease of the respiratory system. In that situation, J45 (the ICD or disease code for asthma) is also noted down as a secondary complication in the diagnoses list. Similarly, the patient may have a history of tobacco use that may be of interest for doctors and may also be a requirement for clinical coding, so the appropriate ICD code – Z87 is recorded to represent the patient’s history of tobacco use. Now, these diagnosis codes represent the health status of the individual patient at that time of admission. Assume that the patient is treated well and is discharged after few days; however, after a few months, he has cardiovascular problems and is admitted for angina or chest pain. As a principal diagnosis, I20 is recorded. Upon further medical testing, he is also diagnosed with type-2 diabetes, so ICD code E11 is registered into his medical history for that admission. The earlier history of COPD (because it is chronic) and the person’s history of tobacco use may also be recorded during this admission as well. Thus, we can see that during the admissions, the patient’s comorbidities are recorded along with other important information. These leave a trace of their medical conditions over time like a trail of breadcrumbs. If we connect these dots of information left over time, we obtain a picture of how the individual’s health progressed. This set of connected pieces of information functions as a fundamental basis of an individual’s health trajectory. The trajectory can be conceptualised as a network where nodes represent a disease or condition, and edges between two nodes indicate that those two diseases have occurred.
in subsequent hospital admissions. Figure 3.6 illustrates the concept of an individual’s health trajectory.

Figure 3.6: Health trajectory of an individual chronic disease patient

An individual chronic disease patient’s health trajectory should not necessarily represent all comorbidities of that chronic disease. However, on a large population level, if we merge individual chronic patients’ trajectories by adding up edges between same disease pairs, we obtain an aggregated version of the health trajectory network. This summarised version of the network, termed the baseline network in our thesis, should represent the overall health trajectory for chronic disease patients. While developing the method for baseline network creation, we should focus on the following characteristics:

1. The baseline network should represent the overall health trajectory of patients having a particular chronic disease.
2. Nodes should have attributes that will represent the overall prevalence of diseases.

3. The transition between diseases in subsequent admissions should be reflected by the edges connecting the nodes. Edges should also denote the number of times one disease progressed into another.

4. The time duration for progression from one disease to another will be explored in our predictive model, so time information should be recorded in the baseline network.

3.6.5 Risk assessment through network similarity

Chronic diseases are of comorbid nature. In many cases, chronic disease patients are not aware of their condition because of irregular medical check-ups or lack of specific symptoms. This applies in particular to our context of T2D. Patients may have T2D for years without being aware of it. However, as chronic diseases (including T2D) have other comorbidities, these should be reflected in the records taken during hospital admissions before the admission when the first T2D diagnosis is made. By looking at the traces of comorbidities in records taken during admissions, we can observe the trajectory that the patients are on. Therefore, our assumption is that if we can match the trajectory of a group of non-chronic patients with that of the baseline network, we might be able to predict that group’s risk of developing the chronic disease (i.e., T2D). If a patient is at high risk, it is likely they will stay on the same pathway in the baseline network. The match score or the risk output can be an effective way to identify and rank high-risk patients without any active medical surveillance over large-scale populations. Subsequently, the high-risk cohort can be advised to take the appropriate preventive measures or offered further clinical diagnosis to minimise risk.

The task of risk assessment by comparing two networks (baseline and health trajectory of a non-chronic patient) is an important focus and contribution of our framework. Because this task requires comparing two networks for similarity, we use the social network and graph theory measures. The disease network that we are discussing has properties of a
social network, as well as time-related features. Therefore, we should focus on the following characteristics while comparing networks:

1. If the patient has a history of diseases that have a high prevalence in the baseline network, they receive a higher risk score.
2. If the patient meets particular demographic criteria (e.g., age) that are at high risk of chronic disease, they receive a higher risk score.
3. If the patient has diseases that are also present and possibly occur together in the baseline network, they receive a higher risk score.
4. If the patient’s transition time between diseases largely coincides with the transition time of the same diseases in baseline network, they receive a higher risk score.

### 3.6.6 Adjusting for attribution effect

Disease diagnosis is essentially a logical reasoning process conducted by physicians. In general, it is done by matching or associating symptoms and physiological conditions with known diseases. It is often undertaken as a top-down approach, so when multiple diseases match the presenting symptoms, physicians look for other symptoms or prescribe various tests that eventually cross out the possibilities of alternate diseases and lead to the correct one. Therefore, it is essential to understand which symptoms are associated with which diseases. Logically, this can be thought as a many-to-many relationship between the two sets of symptoms and diseases. In a generic sense, one disease may have one or more symptoms, and one symptom can be associated with multiple diseases. To uniquely diagnose a disease, physicians need to find symptoms that are exclusive to the particular disease in question. This overall concept can be termed *attribution*.

The original attribution theory is credited to psychologist Fritz Heider (Sanderson, 2009, p. 112) after his phenomenal work in social psychology (Heider, 1944, Heider and Simmel, 1944) in the early twentieth century. The concept had been adopted in different fields of
sciences, including finance and medicine (Kessler, et al., 1999, Shiller, 2003). In general, attribution theory focuses on understanding the factors that are responsible for a particular event. When the factors are properly understood, they are used to predict future occurrences of that particular event. In medicine, especially in medical diagnostics, attribution tries to make inferences about the factors responsible for a particular medical condition (i.e., behaviour or behavioural outcome). The factors can be both internal (e.g., pathophysiological) or external (e.g., environmental). In our framework, we will consider the attribution effects of different comorbidities (i.e., diseases that occur together) that lead to a particular chronic disease. With the aid of network theories, the proposed methods in the framework will give more weight to the diseases that have higher attribution towards the particular chronic disease. The detailed methods of calculating the attribution effect within our predictive framework are discussed in the next chapter (Section 4.8). We further note that, in our empirical analysis of the framework, we also experimented with the advantages of the attribution effect. That is to say, we ran the analysis twice, with and without considering the attribution effect, and compared the performance of both variations.

3.7 Functional Components of the Framework

In the preceding sections, we have discussed the major theoretical concepts that form the basis of our framework. We now discuss the functional components of the framework that are derived from the aforementioned concepts and how they fit together in our research. This discussion is brief, as further details of the methodological components of the framework are discussed in the next chapter.

The input to the research framework is an administrative health dataset. It is assumed that for each patient this should at least contain disease codes representing the health condition of the patient at a point of time (e.g., during hospital admission), along with vital demographic information like age, sex and smoking status. The output of the framework takes two forms: first, it creates a baseline network, which represents an aggregated health trajectory for all chronic disease patients. Second, based on the baseline network,
the framework should provide an interface for predicting the likelihood of chronic disease development for a new test patient who is as yet non-chronic. The interpretation and utility of the baseline network and its predictive capability will depend on the dataset, context and interests of the stakeholders. To illustrate this capability, we implement the framework empirically on an administrative dataset in Chapter 5 and provide the interpretation based on the context.

Figure 3.7: Functional components of the framework

Figure 3.7 shows the functional components of the framework. The overall framework has two major parts, each of which offers an output, as described in the previous paragraph. These two outputs are discussed below.
3.7.1 Part 1: Generating a baseline network

The first part of the framework focuses on generating the baseline network (concept introduced in Section 3.6.4). This network should represent the health trajectory of particular chronic disease (i.e., type 2 diabetes or T2D) patients. Administrative dataset containing medical histories of patients serve as the input for this part. Initially, this dataset goes through the stages of performing an integrity check, code translation and organisation of the data. This, in combination, can be treated as data filtering. The resultant dataset is then ready for analysis. We then need to partition the data into chronic and non-chronic cohorts. Depending on the context and requirements, one chronic disease is selected—for example, T2D—before implementing the framework. Then, if a patient in the research dataset has T2D, he is placed in the chronic cohort; otherwise, he is moved to the non-chronic cohort. Further, each of the cohorts should be partitioned into two groups: the baseline and training partitions. As the names suggest, the baseline partition is used to construct the baseline network, and the other partition is reserved for training the predictive model later.

All medical histories of patients up until they are first diagnosed with the chronic disease (i.e., T2D) in the baseline partition are iteratively aggregated to form the first baseline network, named the ‘positive baseline network’. This network represents the disease progression (i.e., patterns, prevalence) that lead to the particular chronic disease. Similarly, a ‘negative baseline network’ is derived from the baseline partition of the non-chronic patients. The process of generating a baseline network from multiple patients’ medical histories is done through Statistical Aggregation, discussed in detail in the next chapter (Section 4.7). Finally, the two baseline networks are merged into one network by adjusting for the attribution effect, as discussed earlier (see Section 3.6.6). In this process, different properties of the baseline network—for example, prevalence of disease and transitions—are given higher scores if they are more exclusive in the ‘positive baseline network’ and are given lower (negative) scores if they are more exclusive in the ‘negative network’.
baseline network’. The resultant output is the simply called the *baseline network* and represents the final output from the first part of the framework.

**3.7.2 Part 2: Generating predictive model**

The second part of the framework focuses on generating a predictive model for the chronic disease of interest. Similar to any generic predictive model, this part is divided into training and test phases. In general, the predictive model compares the baseline network with the medical history of a patient who is not yet diagnosed with the particular chronic disease, and calculates this patient’s risk of developing the chronic disease in future. The comparison method used here is called *Longitudinal Distance Matching*, which calculates how similar the test patient’s network is, compared to the baseline network. The similarity is scored against three risk factors based on the graph and social network theories. Detail formulation of these scores is given in the next chapter – “Methodology and Framework”.

Along with these scores, which are calculated by comparing with the baseline network, the framework should also consider demographic and behavioural risk factors of the test patients, such as age, sex or smoking history. Therefore, there are multiple risk scores for one test patient and the overall risk of chronic disease is calculated by developing a linear prediction model from these scores. In this process, weighting factors (i.e., parameters) are associated with each of the risk factors. The optimal values of the weighting factors are calculated through standard binary logistic regression, or a parameter estimation model. Optimal values for these parameters should be determined using the training partitions from each cohort (established in the previous step). The framework can now be able to predict a test patient’s likelihood of developing chronic disease by comparing their medical history with the baseline network and calculating the risk scores, as the weighting parameters are known and optimised from the previous training phase. We should also consider data mining approach for prediction. For example, binary classification method can be used. The choice of predictive model can differ based on the context and performance. It should also be interesting to see which predictive modelling gives better
performance. Therefore, in our implementation of the framework, we follow an exploratory approach, running different models and discussing their relative performance.

3.8 Summary

In this chapter, we have introduced the dataset and functional concepts of the framework in the context of Australian health system. After Chapter 2 established the methodological background in an abstract way in relation to the current literature on disease prediction, it was necessary for this chapter to discuss the specific characteristics of a real-world dataset and other contextual information. This included a detailed description of how the data is generated and transmitted, enabling a closer look at common inconsistencies that we should be aware of. We then discussed associated security and privacy issues. Then we introduced the database structure that is suitable to support our methodological framework. Finally, we discussed some key concepts related to the framework and then discussed how these concepts fit together to make up the key components of the framework.

Having introduced the contextual background of the administrative dataset and the framework, we now focus on the methodological and mathematical details of the framework. We will discuss these in the next chapter.
Chapter 4: Methodology and Framework

In this chapter, we introduce our methodology and framework in detail. We also focus on the formulation of methods using mathematical notations and algorithms when possible. We begin by providing an overview of the whole framework. After that, we formalise the conceptual definitions used in the framework. Then we explain the methods involved with each of the components as well as the relations between them. We also discuss the process for measuring performance and validate the framework.

4.1 Overview

We divide our research framework into two major steps. In the first step, we generate the baseline network. This network will represent the health trajectory of patients who have a certain chronic disease in common; for our research context, this will be type 2 diabetes (T2D). The baseline network is derived by analysing all the T2D and non-T2D patients’ medical histories over the full period. The process starts with taking the patients’ medical histories from the healthcare database and ends by generating the baseline network. Two major intermediary steps, namely Patient Filtering and Statistical Aggregation, are involved in generating the baseline network; each of these is discussed in detail later on.

The second step compares the baseline network with the medical history of a patient who is not yet diagnosed with the chronic disease and calculates that patient’s risk of developing the chronic disease in the future. The comparison method used here is called Longitudinal Distance Matching, which calculates how similar the patient’s network is in relation to the baseline network. The calculated similarity score will indicate the risk of developing the chronic disease. Figure 4.1 shows the summarised framework with all the major intermediate steps.
Before going into the details of the framework, we should clarify some of the social network and graph theory-based definitions used in the framework. These conceptual definitions will act as a basis of some of the methods employed. In the next section, we introduce these measures and their definitions, and provide mathematical representations to quantify them.
4.2 Graph Theory-Based Definitions Used in the Framework

We use several theories and concepts from graph theory in order to formalise the research goal and framework. These definitions will constitute the building blocks of our framework, and will also be used later on as a tool for visualisation. In the following subsections, we introduce the graph theory-based definitions.

4.2.1 Individual disease network

A disease network for an individual patient represents the health trajectory of that patient. The health trajectory shows the patient’s transition from one disease to another during subsequent admissions in a healthcare environment. It also indicates the transition of the patient from one point of time to another. We already discussed in the previous chapter (section 3.6.4) about the concept of health trajectory. We also described how to represent an individual’s health trajectory by constructing a disease network. We now define it in terms of mathematical notation.

The nodes of the individual disease network consist of diseases that are present in any of the admissions for that patient. The relation between any two nodes indicates that the patient has progressed from the one disease to another in two subsequent admissions. The relation is directional: that is, the disease that the patient has in the former admission is the source node, and the disease that he has progressed into in latter admission is considered the target node. If there are multiple diseases in any admission, then we consider all possible disease pairs. Also, when the patient has multiple diseases recorded in the same admission, the relations are shown as bi-directional edges between possible disease pairs of the same admission.

To formalise the notations for the above described concepts, suppose we want to construct the disease network for a patient \( P \) given the following properties:
• $P$ has total $n$ hospital admissions denoted as $a_1, a_2, ..., a_n$. The subscripts increase in chronological order in terms of the date of admission. Therefore, $date of admission_{a_i} > date of admission_{a_j}$, when $i > j$.
• $m$ is the total number of possible diseases.
• $D = \{d_1, d_2, ..., d_m\}$ is the overall set of diseases that are being considered.
• $D_{a_i}$ is the set of diseases or comorbidities that are recorded during admission $a_i$ such that $1 \leq i \leq n$.
• $\forall i (D_{a_i} \subset D)$, i.e., for all admissions, recorded comorbidity set $D_{a_i}$ is a subset of overall comorbidity list $D$. In other words, each disease or comorbidity in $D_{a_i}$ must be present in $D$.

Therefore, the disease network for patient $P$ is $N_p(V, E)$, where $V$ is the set of nodes or vertices and $E$ is the set of edges between any two nodes, such that:

• $V = \{v | v \in D_{a_i}, v \in D, 1 \leq i \leq n\}$, i.e., every node $v$ of the disease network $N_p$ must be present in any of the admissions’ comorbidity lists for that patient.
• $E = \{(v_1, v_2) | v_1 \in D_{a_i}, v_2 \in D_{a_j}, 1 \leq i \leq j, v_1 \neq v_2 \text{ when } i = j\}$, i.e., an edge is an ordered set of two nodes that are present in the same or subsequent admissions where the source node is present in the former and the target node is in the latter. In cases when source and target nodes are the same in consecutive admissions, the edge will essentially be a self-loop. It should be noted that there will be edges between all possible disease pairs (except in self-looping cases) that are recorded in the same admission.
• Bounds for $V$ are $1 \leq |V| \leq m$, as $m$ is the overall number of possible diseases. In practice, a patient may not have any admission information, and therefore may not have any recorded diseases. However, we do not include such patients in our calculations, as they do not contribute sufficient information to the framework. Accordingly, the lower bound of the node count is 1: that is, a patient may have only one admission with single recorded disease.
• Bounds for $E$ are $0 \leq |E| \leq m^2$. If only one admission is present, no edge can be formed. There is theoretically no restriction on how many admissions a patient can have; however, as there can be maximum $m$ diseases in the nodes, therefore, these $m$ nodes can form $m^2$ edges among themselves. Allowing for the provision of the self-loop, the maximum possible edges in a directed graph of $m$ nodes is $m^2$, which is the upper bound for this case.

4.2.2 Attributes of individual disease networks

Following up with the earlier section, which discusses formalising individual disease networks, we now discuss some of the associated network attributes. Essentially, we use three attributes: one node-level and another two edge-level attributes. The node level attribute is *node frequency*, which is associated with each disease in the node list (i.e., $V(N_P)$) of the patient’s network $N_P$. This attribute indicates how many times that particular disease has been recorded in the patient’s admission history. The lower bound for node frequency is 1, as the frequency is associated with graph nodes, and graph nodes will exist only if the corresponding disease is recorded at least once in admission history. On the other hand, a disease can be registered in all the admission records for a patient. Therefore, the upper bound for node frequency is equal to the number of admissions.

Each edge of the disease network also has a *frequency* attribute similar to that of node frequency. *Edge frequency* for an edge from disease $d_1$ to $d_2$ represents the number of times a patient has progressed from disease $d_1$ to disease $d_2$. The lower bound for edge frequency is 1. The upper bound for edge frequency is reached when a disease pair is present in each of the admissions. In that case, the frequency is counted once for each admission and once for all the subsequent admissions. If there are $n$ admissions, there are $n - 1$ consecutive admissions. Therefore, given $n$ is the number of admissions, the bounds for edge frequency are:

$$1 \leq \text{edge frequency} \leq 2n - 1$$
The third attribute is also associated with each edge, and represents the time information related to the edge. An edge connects two diseases, indicating that these two diseases have occurred in the same or consecutive admissions, while the time attribute shows the duration of that progression between two admissions. This time attribute has two parts: average time delay and standard deviation. Both are statistical measures of time information. As a particular progression between two diseases can occur multiple times, thereby increasing the other node attribute (edge frequency), we take the duration (admission time) for each of the similar progressions and calculate the statistical average and standard deviation. Collectively these two measures constitute the edge attribute. Figure 4.2 shows the process of constructing an individual disease network and its attributes.
4.2.3 Baseline Network

A baseline network represents the overall health trajectory of a population. Similarly to individual disease networks, the baseline network reveals the disease progression or health trajectory for the population as a whole.

The baseline network is generated from the individual disease networks of patients who constitute the overall population of the baseline network. Individual disease networks are statistically aggregated, and the attributes are updated iteratively to form the baseline network. The process of the aggregation is discussed in detail later in this chapter (see Section 4.7).

4.2.4 Graph attributes of the Baseline Network

The baseline network has similar attributes to the individual disease networks. Instead of revealing the health status of an individual, they reveal health status at a population level. The node frequency attribute shows the number of times the associated disease has occurred in a population during their admissions, while the edge frequency attribute indicates the number of times people have progressed between the diseases associated with the source and target nodes of that particular edge. The time duration attribute reveals the longitudinal nature of the progression: that is, the average duration of the progression and the standard deviation measure the spread of the time duration between that particular pair of diseases in the whole population.

4.3 SNA-Based Definition Used in the Framework

The baseline network, along with its attributes, reveals the health trajectory of a population of chronic disease patients. Social network and graph theories provide several measures that can reveal underlying dynamics and structure of the network that may be otherwise hidden. For our framework, we use some of the node-level measures. These measures are discussed below.
4.3.1 Degree, in-degree and out-degree centrality

Degree centrality indicates activity or prominence. It is a node-level measure, meaning that we can calculate the centrality measures for each node. However, the definition can be extended to the context of the whole network, and therefore it can be used as a network-level centrality measure as well. For the case of a simple graph (i.e., unweighted and undirected), the degree centrality of a node shows the number of neighbouring nodes that are directly connected with the node in question. This essentially means the number of edges that the node has: if a node has a high number of edges, this indicates that the node is more connected in the network, and therefore bears a high degree centrality. Thus, the degree centrality measure indicates the activeness or participation strength of a node.

We need to extend the concept of degree centrality of a simple graph towards our baseline network, which is directed and weighted. The baseline network may also contain self-loops. For directed networks, there are two additional variants of degree centrality, known as in-degree and out-degree centrality. The in-degree centrality of a node shows the number of neighbours that directly connect to that node via directional edges: that is, in-degree centrality is the sum of inbound edges. Similarly, out-degree centrality is the count of outbound edges. If a node has high in-degree centrality, that means more nodes in the network are choosing it. Therefore, in-degree centrality is a measure of popularity in the network, especially applicable to social networks. Similarly, a high out-degree centrality indicates that the node is choosing or creating bonds with more neighbours. Therefore, out-degree is a measure of activeness. The degree centrality measure dictates the overall activity or prominence of a node by summing up its in-degree and out-degree centrality measures. Finally, for the baseline network of diseases, the edges have associated weight in terms of edge frequency. So, for the in-degree centrality of a node, we sum up the edge frequency attributes that are incident to that node instead of just counting the number of incident edges. Similarly, the out-degree centrality of a node is calculated by summing up the edge frequencies of outgoing edges from that particular
node. As before, overall degree centrality is the sum of in-degree and out-degree centrality.

Mathematically, we can formulate the definitions of degree-related centralities as follows:

- Suppose a baseline network \( N_B \) or \( N_B(D, E) \), with \( D \) as set of nodes and \( E \) as set of edges.
- \( N_B \) has \( m \) nodes, denoted as \( d_1, d_2, ..., d_m \).
- \( E \) is the set of edges.
- \( E_{dx,dy} \) is any edge from node \( d_x \) to \( d_y \), where \( E_{dx,dy} \in E \).
- Edge frequency of \( E_{dx,dy} \) is denoted by the function \( freq(E_{dx,dy}) \).

Therefore, the in-degree centrality of node \( d_i \) is \( C_{Di}^{in} \) and calculated as

\[
C_{Di}^{in} = \sum freq(E_{dx,di}), \text{ where } E_{dx,di} \in E
\]

Similarly, out-degree centrality of node \( d_i \) is \( C_{Di}^{out} \) and calculated as

\[
C_{Di}^{out} = \sum freq(E_{di,dx}), \text{ where } E_{di,dx} \in E
\]

Finally, the overall degree centrality of node \( d_i \) is \( C_{Di} \) and calculated as

\[
C_{Di} = C_{Di}^{in} + C_{Di}^{out}
= \sum freq(E_{dx,di}) + \sum freq(E_{di,dx})
\]

Figure 4.3 shows an example of calculations for in-degree, out-degree and degree centrality in the baseline network.
4.3.2 Disease cluster

A cluster within a social network is a group of actors who share lots of interaction with each other, and relatively fewer interactions with the actors outside of the cluster. Actors within the same cluster may indicate that they are functionally similar. This is particularly true for social networks, where relations (e.g., friendship, kinship, gene-association) follow the homophily or ‘birds of a feather’ principle, and similar actors tend to form more links with each other. Similarly, as discussed in Chapter 2, diseases do not occur in isolation, and where diseases share similar pathophysiological, environmental or behavioural risk factors, they tend to occur together. Therefore, the presence of edges between nodes and their edge frequencies are not expected to be random or uniformly distributed in the baseline network. Rather, comorbid diseases are expected to be closer and have more connections between them; thus, they may form clusters. For that reason, in our baseline network for T2D, we focus on looking at the clusters of diseases that frequently occur together.
The process of identifying clusters in networks is done by a community detection algorithm. The implementation method varies substantially between algorithms. In one such method, one group of algorithms tries to detect clusters by first dividing the whole network into different sub-networks (either arbitrarily or a preset number) by removing (cutting) certain edges. Then the algorithm sums up the edge frequencies of the edges that are removed (i.e., the edges that lie between different sub-networks). The algorithm then repeats the process by eliminating a different set of edges and calculates the total edge frequency for inter-network edges. The process continues until the minimum sum of edge frequencies is found, and each sub-network is treated as one cluster. This method is called the minimum-cut method, as it finds the minimum sum of edge frequencies that are cut in order to partition the network into clusters.

One of the challenges of cluster detection is that, if the network is too large (which is often the case), finding the optimum solution deterministically is often exhaustive and computationally expensive. Therefore, most community detection methods apply ‘heuristics’, which means that they relax some of the restrictions of the problems and apply some assumptions. In most cases, this gives a much faster performance, at the cost of some accuracy, which is reasonable in many scenarios. The method we use in this research is also based on heuristics, and the algorithm for implementation is proposed by Blondel et al. (2008). The method uses an objective function (i.e., modularity) that is attempted to be maximised iteratively. The baseline network is partitioned into different clusters and the modularity is calculated for the partition. Modularity is the measure of the density of links inside communities as compared to links between communities (Newman and Girvan, 2004), giving a scalar value between -1 and 1. The network is transformed into an undirected network by ignoring the directional property of the edges. If a pair of nodes has edges in both directions, their frequency attributes are added up and a single undirected edge is considered instead of two directed edges between the nodes (i.e., diseases).
For the baseline network $N_B$, modularity $Q$ is calculated according to the definition of Newman (2004) as follows:

$$Q = \frac{1}{2m} \sum_{i,j} \left[ \text{freq} \left( E_{d_i,d_j} \right) - \frac{k_{d_i} k_{d_j}}{2m} \right] \delta(c_{d_i}, c_{d_j})$$

where $\text{freq} \left( E_{d_i,d_j} \right)$ is the undirected edge frequency between $d_i$ and $d_j$, $k_{d_i} = \sum_j \text{freq} \left( E_{d_i,d_j} \right)$ is the sum of the frequencies of the undirected edges attached to node $d_i$, $c_{d_i}$ is the community to which node $d_i$ is assigned, and the $\delta(x,y)$ is 1 if $x = y$ and 0 otherwise. Finally, $m = \frac{1}{2} \sum_{i,j} \text{freq} \left( E_{d_i,d_j} \right)$. Figure 4.4 shows the concept of clustering in a sample baseline network.

**Figure 4.4: Cluster detection in a sample baseline network**

### 4.4 Database Preparation and Filtering

Administrative datasets are heterogeneous and complex. As the dataset is collected over several years in different healthcare facilities, we should first inspect the attributes and formatting of the dataset, look for inconsistencies and correct them before doing any actual analysis. The dataset should be organised into a structured database that is suitable for our analysis. Further, we need to consider privacy issues, because of the sensitive
nature of the data. In this regard, before beginning the actual analysis, we need to inspect thoroughly, sanitise and organise the dataset, which involves different steps. This section discusses the main components of the database preparation step.

4.4.1 De-identification

Healthcare datasets may contain sensitive information, including social security or Medicare number, patient’s name, home address and exact date of birth. As a result, before beginning the analysis itself, we should take adequate measures to ensure that the data is properly sanitised. For most of the part, our research framework does not require any such information that may be deemed sensitive in nature (e.g., patient’s name, address or Medicare number); the only exception is that we need the date of birth to calculate the patient’s age, as it is an important risk factor for most chronic diseases. However, we do not need to know ages precisely to the day, so we can remove the day and month information from the date of birth, keeping the birth year only. This will significantly reduce risks of re-identification. Next, names and postcodes are stripped from the dataset, as they are not required. Names are, however, replaced by random and uniquely generated IDs to identify patients, as data about the same patients is spread over several database entities. While removing personal information, one should also consider provisions for linking these dispersed records (discussed previously in Section 3.4.1) with other data obtained from different sources. If the dataset will have possible linkage requirements, personal identification should be removed in such a way that it can be linked with other datasets later on. For our framework, we do not have any linkage requirements; therefore, this issue did not arise.

4.4.2 Database organisation

After the data is properly sanitised, the framework focuses on the proper organisation and structure of the data. In most cases, healthcare data comes in an electronic format (e.g., database), and normally the dataset is structured following the owner organisation’s standard. However, it may not still be enough to apply our methods readily on the dataset
before doing some preliminary reorganisation. One potential set of problems with the received dataset is that of integrity. For example, same data may be present in multiple records over different data tables. One possible reason for this duplication is that hospital data can be recorded in several stages. A patient file is maintained during the hospital stay, where different doctors can input their detailed diagnostic and clinical notes; doctors may also refer to medical tests, whose reports are also attached in the file. The patient file is often maintained in paper format until a clinical coder prepares the HCP data from it. The billing department also keeps track of the medical items against which the patient or the insurers are billed. The health insurers, therefore, may receive the data in two different forms: the patient’s HCP data, and the claim data for the same patient from the billing department. These two forms may both be transmitted to the researchers, incurring data duplication.

Another potential problem with a dataset can be discrepancy: that is, part of the admission information for a patient may be available in one record, and another part of the same admission information may be present in another record in the received data, depending on the ways in which they were recorded and structured. Therefore, if we were to keep the original database structure intact, we would need to run the methods on multiple tables to obtain the full set of information. Therefore, we should organise the dataset before doing the analysis and ensure that the data integrity is ensured.

The SQL-based relational data structure is adequate for the framework. The design of the database should firmly implement the primary key–foreign key relationship to ensure consistency. For example, we know that a patient can have multiple admissions. On the other hand, each admission should be associated with exactly one patient. Also, no two patients can have the same identifier. These relationships can easily be implemented in SQL-based database designs by making the patient identifier in the patient table the primary key, and creating a foreign key in the admission table that must point to the primary key. In this way, there would be no possibility, even accidentally, of breaching the integrity of the patient–admission relation. Further, the main entities—patient, provider,
admission, treatment and DRG—should be logically separated in the database by putting them into different separate tables. This database design is also capable of applying complex queries, either directly in the database client engine or through the framework. As part of the analysis methods, we often need to perform such queries in order to run some complex analyses, and the relational design will significantly improve performance, and moreover ensure data security.

From the above discussion, we can discern clear reasons why it can be necessary to reformat or restructure the original dataset. We have already discussed the logical database structure for the framework in the previous chapter (see Section 3.5) in great detail. Therefore, in the next section, we move on to discuss assessing the data.

4.4.3 Preliminary assessment

This is the final step in a three-step process of preparing the dataset. In this step we focus on assessing the coding quality present in the data and, based on that assessment, removing any records that do not have sufficient information to be considered in the analysis. This will ensure that the data is noise-free to the greatest possible extent, which can otherwise affect the overall performance and accuracy of the framework. In assessing the data, we will look for several data characteristics across the entities. These characteristics are given as follows:

- All patients should have sufficient duration of time represented in their records over which we can trace their admission histories. If we include patients with insufficient time information, it may introduce outliers and noise in the disease progression network. The exact upper and lower bounds of time duration depend on the quality and nature of the dataset, as well as the pathophysiology of the specific chronic disease that the framework is analysing. Data sourced from health insurers can include the patient’s joining date and termination date (the lack of which may indicate that they are still members); these can be used to calculate the duration of a patient’s record, as we can assume that during that time all
admission information was sent to the insurers. Sometimes joining and termination dates can be omitted or obfuscated for privacy reasons. In that case, we can consider the duration between first and last admissions. For the part of the framework where we construct the baseline network from chronic disease (i.e., T2D) patients, we examine the period of their record up to the point when they are first diagnosed. Therefore, the effective duration is calculated from the joining date (if available) or first admission date until the admission date at which the first chronic disease code appeared.

- Some patients may have medical conditions that require recurrent admissions. For example, a patient may have a physical injury requiring frequent admissions for dressing. Alternatively, a patient may need regular medical services (e.g., dialysis). As a result, their comorbidity information will be recorded each time they are admitted. Allowing these records can lead to overestimation of their comorbidities, and thus may introduce bias. Therefore, we should set a threshold that allows patients a certain maximum number of admissions per year.

- There should also be a minimum threshold in terms of admission numbers to be considered in the framework. As the framework initially constructs a disease progression network for an individual, having a very small number of admissions will not give reliable information on how the diseases progressed. Also, theoretically we need at least two admissions for each patient, each of which should have valid disease codes, in order to depict transitions over time.

- Some diagnoses or treatment codes that do not affect the chronic disease progression or onset should be excluded from the framework. For example, accidental or physical injuries that are not related in any way to the chronic disease should be omitted. Codes related to consultation with a GP and general diagnostic tests (e.g., full blood count) do not carry significance for our task of predicting chronic disease. In addition, some specific physical conditions or attributes present in the admission record may not make sense if considered alone. Therefore, those conditions or attributes should also be ignored from the analysis. For example, if
conditions like fever, vomiting or vertigo are present during admission, they are recorded in the HCP data; however, these conditions do not contribute to the framework, and hence are put in the exclusion list.

4.5 Disease Code Grouping

The framework uses disease codes, typically in the form of ICD codes, as the core data items. A typical administrative dataset of moderate size should have a large number of these disease codes. For example, in our implementation of the framework we used the ICD-10-AM version of the codes; this version has approximately 20,000 unique and active codes (ACCD, 2015), many of which are likely to be present in the dataset. It would be difficult to run analyses and comprehend the results if we considered all codes individually. Therefore, we need to group the disease codes to shorten the overall number of nodes in the baseline network. There are two ways of accomplishing this, and we implemented both of them in the analysis. A short summary of each method is given below.

4.5.1 ICD code collapse

ICD codes, as well as other clinical codes, follow a hierarchical structure for logical classification, readability and ease of use for both humans and software. The codes have a main segment that refers to a disease group or similar complication. Optional suffixes can be added to reveal further information about the diseases, such as their aetiology, anatomical site or severity. For example, disease codes E08.311 and E08.319 both refer to diabetes mellitus due to an underlying condition with unspecified diabetic retinopathy. The only difference is that the former is related to symptoms with macular oedema, and the latter is linked to symptoms without macular oedema. These differentiations are mostly done for the physician’s reference, record keeping and billing purposes, and our calculation does not require the level of specificity represented by the longer ICD codes. Besides, including them may induce a false sense of community structure (i.e., cluster) in the baseline network, because codes that are derived from the same core ICD are already
related by their classification, and are expected to have close ties with each other, thereby forming clusters. These pseudo-clusters may undermine the actual clusters of interest that may reveal useful information. Moreover, including deeper levels of ICD codes could shift prevalences (i.e., frequency distributions) from the core ICD groups to themselves, and thus may not give a proper picture of prevalence. Therefore, we need to trim the trailing suffixes of the ICD codes that provide only supplementary information about the core disease. For our framework, we only keep the first three characters of the ICD codes during calculation.

4.5.2 Selective comorbidities

In this method, we first select a group of comorbidities, and each node of the baseline network corresponds to one comorbidity. A comorbidity is essentially a generic disease or health condition (e.g., type 2 diabetes, obesity). The comorbidities are selected in such a way that only the relevant diseases or health conditions are included. For example, in our research we are focused on chronic diseases, specifically type 2 diabetes. So, we should choose those comorbidities or health conditions that are likely to be relevant and possibly occur together with diabetes. There are already well-established comorbidity indices that provide related lists. For our research, we adopt the Elixhauser index, one of the most common indices. Our framework uses this index at its base, and adds a few more comorbidities or health condition that seem important according to the context. We have already discussed the Elixhauser and other comorbidity indices in Chapter 3 (see Section 3.6.2).

4.6 Cohort Selection

After preparing the dataset and filtering out the records that have insufficient information, we now focus on the process of creating baseline network. Figure 4.5 shows the steps of the cohort selection process detailed within the overall framework.
In our framework’s application to T2D, there are two cohorts: the cohort of patients diagnosed with T2D, and the cohort of patients who are non-chronic—that is, they do not have T2D. Throughout the analysis, we will refer to the cohort of diabetic patients as $C_{T2D}$ and the cohort of non-diabetic patients as $C'_{T2D}$.

There are well-defined ICD codes for T2D and other chronic diseases. These codes are used to decide whether a patient will belong to the T2D cohort or the non-T2D cohort. To do that, the patients’ diagnosis codes at each admission are compared against the T2D ICD code; if a match is found, the patient is flagged as a member of cohort $C_{T2D}$. Then the date of that admission is saved in the ‘first diagnosed date’ field of the patient table, as that date represents the first time the patient was diagnosed with T2D. If no match is found, the cohort flag is set to indicate that the patient belongs to cohort $C'_{T2D}$ and the ‘first diagnosed date’ is kept blank.

After the patients are divided into two cohorts, they are used to develop a baseline network using their admission history. The partitions within the cohorts are used for validation. Before going into the details of these processes, we need to discuss the process of shortening the ICD codes in order to keep the size of the baseline network within a limit.
4.7 Statistical Aggregation for Baseline Networks

We generate two baseline networks in this step. The first baseline network, referred to as $N_{B+ve}$, is derived from the cohort of the patients diagnosed with T2D. Similarly, the other baseline network, referred to as $N_{B-ve}$, is derived from the cohort of the patients not diagnosed with T2D (i.e., from $C_{T2D}'$). These two baseline networks represent the typical health trajectories of T2D patients and non-T2D patients respectively. As discussed in the definition earlier, the baseline network includes the ICD-10 disease codes as its nodes. An edge between the nodes (disease codes) represents a relation between those two diseases, indicating that these two diseases tend to occur in the same or consecutive admissions. Two edge attributes are associated with each edge: strength and progression delay. The strength attribute denotes the number of times two diseases have occurred simultaneously or in consecutive admissions. The second attribute, progression delay, denotes the average timespan between occurrences of the former disease and the later one.

Generation of any baseline network through the process of statistical aggregation is calculated in two steps. First, all admission histories are inspected for each respective cohort, and the corresponding disease codes are merged into an intermediate time-detailed network. In this network, the edges contain information on all time gaps in disease occurrence between two consecutive admissions. Figure 4.6 shows a typical merging of two patients’ histories into an intermediate time-detailed network. For example, there are two directed edges from $D_1$ to $D_2$, which means that these two diseases have occurred one after another in both ways: that is, $D_1$ has occurred after $D_2$ and $D_2$ has also occurred after $D_1$. $D_2$ has occurred after $D_1$ at time gap 0 (in the same admission for Patient 1) and at $t_2 - t_1$ (in consecutive admissions for Patient 1). Conversely, $D_1$ has occurred before $D_2$ twice: at time gap 0 (in the same admission for Patient 1) and at time gap $t_4 - t_3$ (in consecutive admissions for Patient 2). Note that, when two diseases occur simultaneously in a single admission, this appears in both of the directed edges between them.
Figure 4.6: Construction of intermediate time-detailed network

To generate the intermediate time-detailed graph, an empty graph is created first. Then two sets, $A_c$ and $A_p$, are created to hold the disease codes for the current and previous admissions. These two pointers traverse the entire admission history of a patient, and each time the disease codes of consecutive admissions pointed to by $A_c$ and $A_p$ are calculated to insert the edge information to build the graph. Below is the algorithm used to construct the time-detailed graph.

procedure timeDetailedNetwork (list $p_d$)
    - create empty graph $G$
    - for each patient $p$ in $p_d$
        - create empty set $A_c, A_p$
        - for each admission $a$ from second to last admission of $p$
            - assign $A_p$ to disease codes prior to admission of $a$
            - $t_p$ = time of $A_p$
            - assign $A_c$ to disease codes of $a$
\[ t_c = \text{time of } A_c \]
- for each disease code pairs \( d_{p1}, d_{p2} \) in \( A_p \)
  - insertEdge \( (G, d_{p1}, d_{p2}, 0) \)
- end for
- for each disease code pairs \( d_{c1}, d_{c2} \) in \( A_c \)
  - insertEdge \( (G, d_{c1}, d_{c2}, 0) \)
- end for
- for each disease code pairs \( d_1, d_2 \) in \( A_p \) and \( A_c \)
  - such that \( d_1 \in A_c \) and \( d_2 \in A_p \)
  - insertEdge \( (G, d_1, d_2, t_c - t_p) \)
- end for
- end for
end procedure

procedure insertEdge (Graph \( G \), Node \( source \), Node \( target \), time \( \Delta t \))
- if \( source \) not exists in \( G \)
  - insert \( source \) in \( G \) with \( strength_{source} = 0 \)
- end if
- if \( target \) not exists in \( G \)
  - insert \( target \) in \( G \) with \( strength_{target} = 0 \)
- end if
- insert edge entry between \( source \) and \( target \) with \( time \text{ gap} = \Delta t \)
- increase strength of \( target \) by 1
end procedure

The second part of the baseline network generation method deals with the merging of the time gap information between the consecutive disease codes of the time-detailed graph. Because of the vast number of occurrences of disease codes all together in the patients, it is impractical to consider all the time gap information contained in the disease codes. Besides, it would also be computationally expensive and would complicate the comparison steps of the next part. To merge all occurrences of consecutive disease codes and corresponding time gap information into a single time duration, we considered the
mean values of time gaps ± one standard deviation. In case the lower bound becomes negative, we set zero as the lower bound. Figure 4.7 illustrates merging time gap information to convert a time-detailed graph into a baseline network.

Figure 4.7: Merging of time gap information to convert a time-detailed graph into a baseline network

Overall, in this step of the framework, we generate two baseline networks. These are referred to as $N_{B+ve}$ and $N_{B-ve}$, the former for the T2D patients and the latter for the non-T2D patients. In the next step, we adjust the parameters of the T2D patients from the $N_{B-ve}$ network using the attribution effect.

### 4.8 Adjustment for Attribution Effect

In general, attribution theory focuses on understanding the factors that are responsible for a particular event. In medicine, especially in medical diagnostics, attribution refers to the process of making inferences about the factors responsible for a particular medical condition (i.e., patient behaviour or behavioural outcome). The attribution principle can be put as follows: if an outcome $M$ is the result of attributes $A$, $B$ and $C$, and an outcome $N$ is the result of attributes $B$, $C$ and $D$, then we can say that $B$ and $C$ both are present and
probably responsible for \( M \) and \( N \). Now, if a patient has attributes \( B \) or \( C \) or both, and we are asked which outcome he or she will have, provided that \( M \) and \( N \) outcomes are exclusive—that is, one cannot have both—then it is difficult, if not impossible, to tell whether \( M \) or \( N \) will occur, because \( B \) and \( C \) are present in both outcomes. However, if the patient also has either attribute \( A \) or \( C \), we can make a better decision, as these attributes are unique to their outcomes. Therefore, to properly predict outcome or risk, we should take the attribution effect in consideration.

The two baseline networks from the previous step create a similar scenario. Suppose, for example, that we find that a disease is more prevalent in positive baseline network, \( N_{B+ve} \), as that disease has higher node frequency; this does not necessarily mean that this disease will be a definite risk factor for chronic disease. Perhaps that particular disease is also prevalent in the non-chronic disease patient cohort, or in the negative baseline network, \( N_{B-ve} \). Therefore, we should not look for more prevalent comorbidities in \( N_{B+ve} \), but rather we should focus on finding prevalent comorbidities in \( N_{B+ve} \) that are also less prevalent in \( N_{B+ve} \). In other words, we should look for more exclusive diseases in \( N_{B+ve} \) by looking at the differences in prevalence, and in the process, adjust for the attribution effect.

To keep the analysis process logically separated, we do not update the attributes in the baseline network for T2D patients. Rather, we generate another instance of a baseline network from \( N_{B+ve} \) and \( N_{B-ve} \). We refer to this network simply as baseline network \( N_B \), which is a composite of \( N_{B+ve} \) and \( N_{B-ve} \) generated through attribute adjustment. The nodes and edges of the new baseline network are essentially the union of the nodes and edges of \( N_{B+ve} \) and \( N_{B-ve} \), which means that the following relations hold true:

\[
\begin{align*}
V(N_B) &= V(N_{B+ve}) \cup V(N_{B-ve}) \\
E(N_B) &= E(N_{B+ve}) \cup E(N_{B-ve})
\end{align*}
\]
Each node of $N_B$, say $d_{iN_B}$, is calculated as the relative difference between that node’s frequencies in $N_{B+ve}$ and in $N_{B-ve}$. There will be two scenarios that we need to consider for this calculation. First, if the node is present in $N_{B-ve}$, then it can be expressed as:

$$freq\left( d_{iN_B} \mid d_{iN_B} \in V(N_{B-ve}) \right) = \begin{cases} \frac{freq(d_{iN_B+ve}) - freq(d_{iN_B-ve})}{freq(d_{iN_B+ve})}, & \text{when } d_{iN_B} \in V(N_{B+ve}) \\ -1, & \text{otherwise} \end{cases}$$

Recall that $V(N_{B-ve})$ is the set of edges (vertices) of the baseline network $N_{B-ve}$. If the node $d_{iN_B}$ is not present in $N_{B-ve}$, the calculation will give a divide-by-zero error (i.e., infinity result). Therefore, if the particular node is absent from $N_{B-ve}$, we assign its frequency in $N_B$ equal to the maximum relative frequency difference calculated for the nodes that are present in both the baseline networks $N_{B-ve}$ and $N_{B+ve}$. Mathematically, the node frequency for such nodes of the baseline network $N_B$ can be defined as follows:

$$freq\left( d_{iN_B} \mid d_{iN_B} \notin V(N_{B-ve}) \right) = \max_{d_{jN_B} \in V(N_{B-ve})} freq(d_{jN_B})$$

After obtaining the baseline network’s node attributes in the above process, the attributes are normalised in the range of 0 to 1 inclusive for analysis and comparison purposes. The process is quite straightforward: each node frequency is divided by the maximum node frequency of the baseline network. Mathematically,

$$freq\left( d_{iN_B} \right) = \frac{\max_{d_{jN_B} \in V(N_B)} freq(d_{jN_B}) - \min_{d_{jN_B} \in V(N_B)} freq(d_{jN_B})}{\max_{d_{jN_B} \in V(N_B)} freq(d_{jN_B}) - \min_{d_{jN_B} \in V(N_B)} freq(d_{jN_B})}$$

In the same manner, each edge of $N_B$ is calculated from the same edges in the T2D and non-T2D baseline networks, considering the attribution effect. The equations and boundaries also remain similar, as above, except that edge notations are used instead of nodes.

Now that we have established the new baseline network, it is considered to represent the trajectory of T2D patients, where the attributes (node, edges, frequency etc.) represent
unique characteristics of progression, as we have considered the attribution effect. In the next step, we will use this baseline network for risk prediction. Unless otherwise specified, in subsequent calculations, we will always refer to this attribute-adjusted network (i.e., $N_B$) when we mention the baseline network. Figure 4.8 illustrates the process of attribution adjustment for the nodes.

Figure 4.8: Example process of attribution adjustment for nodes

4.9 Risk Prediction

This is the second part of our framework, and perhaps more analytically complex than the previous ‘baseline creation’ part. In this phase, we predict the risk of developing the chronic disease for a test patient who has not yet been diagnosed with that disease. To accomplish this, we compare the medical history of that patient with the baseline network of the chronic disease (i.e., T2D) derived in the first part. Note that the baseline network has some important properties, including the number of times the disease has occurred in
all the patients diagnosed with chronic disease, denoted as strength; the average time gap between any disease b’s occurrence after any disease a; and the sequence of disease occurrence, called the direction of the network. When matching the test patient’s network with the baseline network, we consider the following principles:

1. The patient’s chance of developing the chronic disease is increased when he has a history of diseases that have a high strength in baseline network.

2. The patient’s risk is increased if he has more diseases from the baseline network (this principle is analogous with collaborative filtering).

3. The patient’s risk is increased if his sequence of diseases largely matches the sequence of those diseases in baseline network.

4. The patient’s risk is increased if the time gap between any two diseases in the patient matches the time gap of those two diseases in the baseline network.

In addition, we consider behavioural, sex- and age-related risk factors, as these are often associated with the onset of chronic disease. We discuss how to calculate risk score considering these factors in the sections below.

4.9.1 Age, sex and behavioural risk factors

People approaching elderly ages are often at higher risk for chronic diseases, because of various pathophysiological, environmental and lifestyle factors. This is especially true for T2D, as there is strong evidence that age is a risk factor. Some chronic diseases are also likely to have a gender bias. In an attempt to keep the framework generic, we also kept a sex-based risk score for diabetes. Finally, some behavioural risk factors, such as alcohol or smoking, are also considered, as these act as risk factors for T2D. Whether or not the patient has a current or previous history of alcohol or smoking is coded in the HCP data and reflected in the patient records. Therefore, looking at the patient’s recorded list of diagnoses, from the specific ICD codes that represent behavioural risk factors like smoking and alcohol use, we can determine a behavioural risk score for a patient.
The age risk factor \( rf_{age} \) is a continuous score, ranging from 0 to 1. We divide the patient’s actual age (in years) by the difference between the maximum and minimum ages in the cohort in order to normalise the age score within the range 0 to 1. The sex risk factor \( rf_{sex} \) is essentially a categorical score, and needs no further calculation, as the patient record in the dataset already has the flag to indicate whether the patient is male or female. Finally, the score for behavioural risk factors \( rf_{behav} \) has the discrete value of 0 if the patient does not have any ICD codes that are considered as risk scores (e.g., smoking) and 1 if at least one match is found.

After the scores for age, sex and behavioural risk factors are set, the framework can then move forward to calculate the scores for the graph- and social network-based risk factors of a test patient, comparing them with baseline network.

### 4.9.2 Longitudinal Distance Matching

In this part of the prediction framework, we compare the disease network of a test non-chronic patient with the baseline network and give scores against three graph theory and network-based scores. We call the overall comparison method ‘longitudinal distance matching’. The network comparison method and this name are motivated by the concept of the ‘String Edit Distance’ algorithm, also known as ‘Levenshtein distance’ (Levenshtein, 1966). String edit distance methods are widely used in spell checkers, word suggestion and optical character recognition. ‘Edit distance’ is defined as the minimum number of operations (insertions, deletions or substitutions) required to change one word into another. For example, to change ‘worde’ to ‘world’, we need to delete ‘e’ from the end of ‘worde’ and then insert ‘l’ before ‘d’. Thus the cost of this conversion is one deletion and one insertion. This algorithm considers the sequence of the words and calculates the minimum cost of conversion; this is analogous to our target. According to our assumption above, if the baseline network has the disease sequence ‘a-b-c-d’, then a patient’s disease sequence ‘a-b-c’ is more similar to the baseline than another patient’s disease sequence ‘a-c-d’. Although both patient sequences have same number (three) of overlap with the baseline network, the former patient’s edit distance is smaller.
However, for our case, we have two disease networks—the baseline and the test patient’s network—instead of a flat sequence of characters, as in typical string edit distance problems. In addition, the two networks have several attributes that also need to be matched. These factors make the two scenarios different in terms of data structure and implementation. Besides, the edit distance method would be computationally expensive for our case if we were to adapt it for our large sequence of diseases with networked structure. Further, the number of diseases in a test patient’s network should be considerably smaller than the overall baseline network, resulting in a great number of mismatches between the two networks. Therefore, we match the graph similarity against three different network-based risk factors and find the corresponding similarity-matching scores. Scores for each risk factor have a mathematical formulation, and their motivations come from network theory and SNA (see Chapter 2). These scores are discussed below.

**Graph node match score**: This measures the similarity between the test patient network ($N_{test}$) and the baseline network ($N_B$) in terms of disease prevalence. As diseases are depicted as the network nodes, this measure is considered to calculate the ‘node-based risk factor’, or $r_{f_{node}}$. The measure also considers the prevalence intensity, or node frequency, of the baseline and test patient networks while calculating the score. Therefore, for a test patient, a high value in graph node match score is only possible under these following three scenarios:

1. The test patient has more diseases that are also present in baseline network.
2. These diseases (that are common in both networks) have higher prevalence in the baseline network.
3. These diseases (that are common in both networks) have higher prevalence in the test patient’s network.
Mathematically, we can define scores for the node-based risk factors $r_{f_{\text{node}}}$ of any node $d_{\text{test}}$ of a test patient’s disease network $N_{\text{test}}$ as follows:

$$r_{f_{\text{node} d_{\text{test}}}} = \begin{cases} \text{match score} & \text{total preval} \neq 0 \\ \text{total preval} & 0, \text{otherwise} \end{cases}$$

where

$$\text{match score} = \sum_{i=1}^{\left| V(N_{\text{test}}) \right|} \text{freq}(d_{\text{test}}) \ast \text{freq}\left(d_{i_{NB}}\right), \text{where } d_{\text{test}} = d_{i_{NB}}$$

$$\text{total preval} = \sum_{i=1}^{\left| V(N_{\text{test}}) \right|} \text{freq}(d_{\text{test}})$$

The numerator, the ‘match score’, essentially multiplies the frequency of a common disease in both baseline network and test patient’s disease network, and sums the results over all common diseases of both networks. The denominator is used to normalise the score in terms of the overall sum of frequencies for the nodes of the test network. If the denominator is 0, the score is kept at 0 to avoid divide-by-zero error, although the possibility of this does not arise, as we checked earlier during the filtering step.

**Graph pattern match score:** Like the node-based risk factor, the graph pattern match score measures similarity in the disease transition: that is, the calculation considers the number of matching edges and their corresponding frequencies. The risk factor against which the score is given is denoted as the edge-based risk factor, or $r_{f_{\text{edge}}}$. The equations are also similar to those of the graph node match score. The only difference is that, instead of disease prevalence, the calculation involves the transition prevalence between disease pairs; that is, it looks for edge frequency.

**Graph cluster match score:** This measure is scored against the cluster-based risk factor, $r_{f_{\text{cluster}}}$. The process is slightly different from the previous two network-based risk factors. This risk factor is based on the social network theory of clustering. The motivation behind the score is that diseases do not occur in isolation, but rather, a group of diseases tend to occur together: these diseases have a higher number of transitions (i.e., existence
of edges, high frequency) between them and lower number of transitions to diseases of other groups. Therefore, this score measures the proportion of edges in the test patients network that lie inside the same cluster in the baseline network.

To calculate the score, the framework first runs a clustering algorithm (Blondel, et al., 2008) in the baseline network. The algorithm assigns ID numbers to all nodes (diseases) of the network. If two diseases in the network receive the same ID, it indicates that the diseases are in the same cluster. Once the baseline network nodes receive the clustering IDs, the test network is brought to measure the cluster similarity. To do that, each of the edges of the test network is considered. If the nodes that make up the edge have the same clustering ID in the baseline network, we count that as a cluster match. The final match score is the sum of all cluster matches normalised by the number of edges present in the test network. Therefore,

$$\frac{rf_{\text{cluster}}}{\text{count of edges in } N_{test} \text{ whose corresponding nodes have same cluster ID in } N_{B}} = \frac{\text{count of total edges in } N_{B}}{\text{count of total edges in } N_{B}}$$

4.10 Parameter Estimation Model

The framework utilises a parameter estimation model to optimise the weighting factors for the parameters that are generated in the above calculation. In the framework, we have six risk factors as independent variables, which are listed in Table 4.1.

In our current analysis, risk factors for sex were calculated from the literature review on T2D. If the prevalence of chronic disease is higher for females, they are assigned a higher weight (between 0 and 1) and males are assigned a lower weight. The exact assignment of values depends on the relevant findings on the prevalence of the disease in question as reported in the literature.
Table 4.1: List of risk factors considered in the framework

<table>
<thead>
<tr>
<th>Risk factor type</th>
<th>Name</th>
<th>Value type</th>
<th>Notation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Socio-demographic</td>
<td>Age</td>
<td>Continuous</td>
<td>$f_{age}$</td>
</tr>
<tr>
<td></td>
<td>Sex</td>
<td>Discrete</td>
<td>$f_{sex}$</td>
</tr>
<tr>
<td></td>
<td>Behaviour</td>
<td>Discrete</td>
<td>$f_{behav}$</td>
</tr>
<tr>
<td>Graph/health trajectory based</td>
<td>Graph node match</td>
<td>Continuous</td>
<td>$f_{gNode}$</td>
</tr>
<tr>
<td></td>
<td>Graph pattern match</td>
<td>Continuous</td>
<td>$f_{gPat}$</td>
</tr>
<tr>
<td></td>
<td>Network cluster match</td>
<td>Continuous</td>
<td>$f_{cluster}$</td>
</tr>
</tbody>
</table>

The overall set of risk factors is defined as:

$$rf = \{ f_{age}, f_{sex}, f_{behav}, f_{gNode}, f_{gPat}, f_{cluster} \}$$

As discussed in the previous section, we obtain scores against each risk factor in $rf$ for all test patients who are not yet diagnosed with the chronic disease (i.e., T2D). For each risk factor, we define a scaling or weighting factor that normally has a value within the range of -1 to 1 inclusive. This range covers equal magnitude of negative and positive values relative to zero. Now, the overall risk factor for a patient is the weighted sum of the scores: this means that we multiply each of the test patient’s individual risk scores with their corresponding scaling or weighting factors, and then add them together. Therefore, the risk score of a test patient is:

$$S_{test} = \sum_{i=1}^{\vert rf \vert} s_i \times f_i$$

Here, $f_i$ is the $i^{th}$ risk factor (i.e., age, sex, gNode) and $s_i$ is the weighting factor of the corresponding risk factor. $\vert rf \vert$ indicates the size of the set of risk factors in a generic sense; in our present case, it is 6.
Optimum values of the weighting factors for each risk score are calculated from a parameter estimation model by training it on a cohort of test patients with known outcomes in terms of whether or not they develop T2D. The purpose of this estimation is to decide which factors are the better determinants of the chronic disease. In other words, the model looks for the effect of the risk factors on the outcome, which is whether or not the patient will develop T2D. For the process of determining so, the model tries to adjust weighting factors for each risk factor and calculates the accuracy of the framework by applying it on a sample of test patients. The adjustments of the weighting factors are done iteratively, and the combination that gives the maximum accuracy is set to the optimum values. Figure 4.9 illustrates the process of choosing the optimum threshold value and accuracy for a particular combination of weighting factors for a group of test patients whose T2D conditions are known.

Figure 4.9: Process of choosing optimum threshold value and accuracy
At first, the framework selects a group of T2D patients from cohort $C_{T2D}$ and a group of non-T2D patients from cohort $C'_{T2D}$. A detailed process of choosing test patients from cohorts is discussed in the next section. For this part, we focus on the process that the estimation model relies on. Let us assume we choose a test cohort of T2D patients and refer it as $C_{test+ve}$, where $C_{test+ve} \subseteq C_{T2D}$. We also test the cohort of non-T2D patients and refer it as $C_{test-ve}$, where $C_{test-ve} \subseteq C'_{T2D}$ and $|C_{test+ve}| = |C'_{test-ve}|$; that is, the test cohorts are of same size. The framework then incrementally assigns values to each of the weighting factors, starting from -1 up to 1, with an increment size of 0.1. This range and increment size can be changed depending on the context later on. Now, the assignment of different sequential values results in different combinations of weighting factors. We can, in fact, calculate the number of possible combinations using the following formula:

$$Number \ of \ combinations \ = \ (upper \ bound - lower \ bound + 1) \times |sf|$$

For each combination of the weighting factor values, the parameter estimation model computes the optimum accuracy of prediction for the current weighting factor values. This is done first by calculating the actual risk score, which is the sum of risk factor scores multiplied by the current weighting factor values for all the test patients. Then the model iteratively sets and increments a threshold score, and considers a test patient at risk of T2D if his risk score is above or equal to the threshold value. If not, then the test patient is considered non-T2D. Now, as the test patient’s actual T2D status is known, we can easily check whether or not this prediction is correct. Thus we count the percentage of correct predictions made for this threshold value, and then update the threshold by incrementing it to a predefined size. Finally, the threshold value that gives the maximum accuracy is considered as optimal from this particular combination of weighting factors.

As we can recall, the framework has generated different combinations of scale factors, and for each combination it tries to discover the possible accuracy of prediction and the corresponding threshold value for considering a T2D patient. Depending on the increment size and the upper and lower bounds for the weighting factors chosen, it may take some
time to calculate all possible combinations. Within this full execution cycle, the model should find several maxima, and the model remembers the combinations that resulted in those maxima. Now, after the execution is over, the combinations that gave the highest accuracy percentage can be regarded as the optimal values for the weighting factors. Alternatively, the model can start another pass, now focusing on a smaller range around the previously identified maxima. This time, the increment size can be made much smaller in order to search for better precision. Undertaking several passes is important, because if the increment size if the first step is set too large, the model can miss actual optimum values. Therefore, iterative passes with smaller ranges around the maxima identified in the last iteration and smaller increment sizes should give optimum values. Also, as mentioned earlier, when calculating accuracy for each iteration, the model also calculates the threshold value to determine whether a score will be considered T2D positive or T2D negative. If the score is greater than the threshold value, the corresponding patient is considered T2D positive, and T2D negative otherwise. The combination that gives the optimum accuracy threshold for that combination is then set as the optimum threshold, because that particular threshold value is responsible for the accuracy.

Once the optimum weighting factors are determined from the parameter estimation model, the framework is now ready to function. The output of the model takes two forms. First is the baseline network \( N_B \), which can be used to understand the chronic disease (i.e., T2D) progression. The second output is the optimum weighting factors, along with the prediction model presented here that can be used to assess the risk of chronic disease (i.e., T2D) of any new patient. To do that, we calculate the scores described in Section 4.9.2. Individual scores are multiplied by the weighting factors determined by the parameter estimation model, and then summed up to obtain the final score. If the final score is greater than or equal to the threshold value, the patient is considered to have a very high risk. The risk assessment process can be iterated over multiple test patients to make a comparison or to assess who is at higher risk.
Having described the framework and its operation, we now discuss the performance and validation methods for the framework.

**4.11 Performance and Validation**

**4.11.1 k-fold validation**

We employ k-fold validation as a method of cross-validation to test how our model can be scaled or generalised onto another dataset for the same chronic disease without changing any parameters. In k-fold cross-validation, the members of the original cohort are randomly assigned to $k$ equally sized, mutually exclusive partitions (Kohavi, 1995). Out of these $k$ partitions, a single partition is assigned as reserved. Members of this partition are used as the validation data for testing the model, while members in the remaining $k - 1$ partitions are used for training the model. After the training and testing related to this partition assignment is completed, the process of training and testing is repeated for another $k - 1$ times. In each repeat of the process, another unique subsample out of the $k$ partitions is used for testing or validating. Therefore, training and testing are run a total of $k$ times, and each time the testing data are unique, as that single partition is uniquely chosen from $k$ different partitions. The $k$ results from the folds can then be averaged or otherwise combined to produce a single estimation. The advantage of this method over repeated random sub-samplings is that all observations are used for both training and validation, and each observation is used for validation exactly once.

For our framework, we used the commonly followed 10-fold cross-validation (i.e., $k = 10$). We divided the cohorts of T2D and non-T2D patients (i.e., $C_i(T2D)$ and $C'_i(T2D)$) each into 10 partitions of equal size. These partitions are denoted as $C_i(T2D)$ for partitions from the T2D cohort (i.e., $C_{T2D}$) and $C'_i(T2D)$ for partitions from the non-T2D cohort (i.e., $C'_{T2D}$); $i$ is the serial of the cohort.
The following relations hold true for the cohorts and cohort partition membership:

a. \( C_{T2D} = \bigcup_{i=1}^{10} C_i(T2D) \)

b. \( C_i(T2D) \cap C_j(T2D) = \emptyset \), where \( 1 \leq ij \leq 10 \)

c. \( C'_i(T2D) = \bigcup_{i=1}^{10} C'_i(T2D) \)

d. \( C'_i(T2D) \cap C'_j(T2D) = \emptyset \), where \( 1 \leq ij \leq 10 \)

For the k-fold validation (\( k = 10 \)), we generated the baseline networks from the first nine partitions of chronic and non-chronic patients—from \( C_1(T2D) \) to \( C_9(T2D) \) and from \( C'_1(T2D) \) to \( C'_9(T2D) \). The baseline networks are then attribute-adjusted, and then weighting factors are estimated from the same nine cohorts. After that, the model is tested on the remaining T2D and non-T2D partitions—on \( C_{10(T2D)} \) and \( C'_{10(T2D)} \). The process is then repeated by using other cohorts individually (9th, 8th etc. up to 1st) for testing, each time using the remaining cohorts as framework building (i.e., baseline network creation).

### 4.11.2 Precision and recall

Precision and recall are two measures are used to determine how well the model is performing. **Precision** reveals the fraction of the result that is relevant. In contrast, **recall** shows the fraction of relevant results that the model is returning. In our framework, we test it against patients whose chronic conditions are known: that is, we know whether they are T2D positive or T2D negative. Now, after generating the baseline network and finding the weighting factors and threshold values, we can test the framework. If for any test patient the computed score is greater than or equal to the threshold value, we conclude that he is at higher risk, meaning that he is progressing towards becoming T2D positive. If, in reality, the patient is diagnosed with T2D, the prediction is correct, while if the patient does not have T2D in his medical history, then we say that the model has given a false positive result. Conversely, if the prediction model says that the patient is not at risk and the test patient has in fact developed T2D, it means that the model has given a false negative result, which is also not desirable. The last option is that, in both predictions and in reality, the patient is not shown to have T2D. This is another case in which the
prediction is correct. Table 4.2 shows the possible outcomes of the test results and their corresponding success or failure status.

<table>
<thead>
<tr>
<th>Actual patient status</th>
<th>Prediction from framework</th>
<th>Validation</th>
<th>Success/Error</th>
</tr>
</thead>
<tbody>
<tr>
<td>T2D +ve</td>
<td>T2D +ve</td>
<td>True positive</td>
<td>Success</td>
</tr>
<tr>
<td>T2D +ve</td>
<td>T2D -ve</td>
<td>False negative</td>
<td>Error (Type II)</td>
</tr>
<tr>
<td>T2D -ve</td>
<td>T2D +ve</td>
<td>False positive</td>
<td>Error (Type I)</td>
</tr>
<tr>
<td>T2D -ve</td>
<td>T2D -ve</td>
<td>True negative</td>
<td>Success</td>
</tr>
</tbody>
</table>

Precision, therefore, in our case, is an indication of how many patients, out of all the patients who are predicted to be at high risk of T2D by the framework, are actually T2D positive. Thus,

\[
\text{precision} = \frac{\text{true positive}}{\text{true positive} + \text{false positive}}
\]

Inversely, recall is an indication of how many patients, out of all the patients who are actually T2D positive, are correctly identified as at high risk by the framework. Therefore,

\[
\text{recall} = \frac{\text{true positive}}{\text{true positive} + \text{false negative}}
\]

### 4.12 Performance Based on Alternate Implementation Criteria

In the whole framework, there are three criteria that have alternate options for execution. Two of these can be seen in the diagram of the framework (see Figure 4.1 at the beginning of this chapter). The first option is to implement the attribution effect during the baseline creation step. The attribute adjustment can give higher weights to the diseases or progressions that are more exclusive to the positive baseline network of chronic patients than to the negative baseline network of non-chronic patients. Therefore, attribution adjustment is likely to give better prediction performance, and hence we keep it as default execution path. The alternative is to not consider attribute adjustment, but rather to construct a baseline network simply from the chronic patients’ positive baseline network.
The second option for an alternate execution path arises during the disease prediction part of the framework. To train and test the model, we considered three different models. Two of them are linear models, namely the parameter estimation model and the binary logistic regression model. The other is from a data mining-based binary tree classification model. All of these models utilise the same baseline network and risk factors at the backend, but differ in the way the models are executed, as well as their performance.

The third option for alternate execution is based on the disease code grouping. By default, the framework uses the Elixhauser comorbidity index, which is a pre-selected group of relevant diseases derived based on clinical practice. The other alternative is to use collapsed ICD codes to three main characters. However, in that way, the number of diseases in the baseline network would be high, and irrelevant codes may be present in the baseline network.

In our implementation of the framework, we are interested in exploring the performance variations between the different alternate methods mentioned above. Therefore, we ran the framework on the same dataset in multiple turns. In each of the turns, we chose different combinations of the alternate options. Then we measured the performance of the framework in terms of accuracy, precision and recall percentage. The performance variations and possible reasons behind them are discussed in the section 5.9 offering insights based on the implementation context.

4.13 Summary

In this chapter, we have discussed in detail the methods that make up the framework. The conceptual definitions of the graph theory-based measures used in the framework were introduced and explained. From there, we discussed different components of the framework in the order they are to be executed on the actual dataset. Finally, we introduced the methods that are to be used in validating the framework. Having explained the framework in full detail, we now move to the next chapter, in which we apply the framework on the healthcare dataset and discuss the results of the implementation.
Chapter 5: Results and Discussions

In this chapter, we present the empirical aspects of our implementation of the framework in a healthcare data context, and discuss our findings from the analysis. The underlying methods for the analysis have already been described in Chapter 4. Contextual specifications about the healthcare dataset we used for analysis have also been introduced in Chapter 3. We begin with the basic statistics of the dataset, followed by the different constraints on the filtering process. Next, we show the results for the various components of our analysis framework. Finally, we present the validation process of the framework.

5.1 Descriptive Statistics

We received the original dataset in the form of relational database records saved in a PostgreSQL database. The full dataset contains hospital admission information for around 749,000 people who were members of a range of private healthcare funds based in Australia. The dataset is essentially a snapshot of these patients’ health status over the years for which they were a member of the health funds and admitted to hospital for treatment within that period. Each patient record also includes basic information such as demographics and joining and termination dates (if applicable) of fund membership. Patients’ admission records are in another linked table: each admission record contains both diagnosis information, in the form of ICD-10-AM, and claim information that was lodged with the insurers, with details of the providers and the services provided. Table 5.1 shows the basic properties and statistics of the original dataset that are of interest.
Table 5.1: Overall statistics of the dataset

<table>
<thead>
<tr>
<th>S/L</th>
<th>Database Table Name</th>
<th>Total Records Imported</th>
<th>Specific Properties or Range</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Members</td>
<td>748,982</td>
<td>Male 339,321 (45.3%)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Female 409,661 (54.7%)</td>
</tr>
<tr>
<td>2</td>
<td>Members having admission</td>
<td>405,428</td>
<td>Male 180,705 (44.57%)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Female 224,723 (55.43%)</td>
</tr>
<tr>
<td>3</td>
<td>Members with diagnosis information</td>
<td>384,538</td>
<td>Male 171,206 (44.52%)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Female 213,332 (55.48%)</td>
</tr>
<tr>
<td>4</td>
<td>Claim</td>
<td>19,118,056</td>
<td>Earliest services date Nov, 1996</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Latest service date Jun, 2015</td>
</tr>
<tr>
<td>5</td>
<td>Admission</td>
<td>1,387,181</td>
<td>Earliest admission date Jun, 2000</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Latest admission date Jun, 2015</td>
</tr>
<tr>
<td>6</td>
<td>Admission diagnosis</td>
<td>9,346,349</td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>Claim metadata</td>
<td>77,224,290</td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>Provider</td>
<td>446,964</td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>Provider Group</td>
<td>1,388</td>
<td></td>
</tr>
</tbody>
</table>

Table 5.1 shows that within the overall dataset population of around 749,000, the proportion of the female population (54.7%) is slightly larger than that of the male proportion (45.3%). Given that in Australia the male-to-female sex ratio at birth is
approximately 105:100 (ABS, 2014), the higher female proportion in the overall dataset may seem a bit biased. However, in reality, this imbalance is possible for several reasons. First, the higher male-to-female ratio in the population at birth decreases with age, because of the higher male mortality rate at younger ages. Second, this effect is further influenced by the working age group (i.e., 15–64 years) through net overseas migration, which historically has seen the departure of more males than females. Further, above age 70, the male mortality rate is significantly higher than the female mortality rate. These facts drag down the male-to-female sex ratio as age group increases. This trend is particularly likely to be represented in our dataset, whose population mostly belong to older age groups, because young people generally do not tend to opt for private health cover. Therefore, as the dataset is dominated by the elderly population, which in reality has a higher proportion of females, the represented higher female ratio of 54.7% seems logical in our context.

Also evident in Table 5.1 is the fact that not all members were admitted within the period that the dataset covered. Only half of the members (54%) were admitted at least once to hospitals. A small proportion (5%) of these admitted members do not have any diagnoses or disease information, leaving around 385,000 members with the valid diagnosis information necessary for inclusion in the framework.

Before moving on to the analysis, we want to look at the admission trends for the whole population to check for any biases or significant trends that we will need to consider in the subsequent filtering process. Figure 5.1 shows the admission trends for the whole population. It includes both the total number of admissions and the number of unique patients accounting for those admissions on a per-month basis. Both trends show a slight increase over the years the data is recorded. The trend lines indicate that in every month, on average, around 153 more admissions took place against the increment of around 97 unique patients. This means that average admissions per patient were increasing over the period. Overall rates of growth in admissions and patients are significantly higher than the population growth rate of Australia, which fluctuated between around 1.3% and 2.3%
(ABS, 2016) during the period. This indicates that more patients were joining the funds and therefore entering the system. Also, the ageing part of the population requires more admissions.

Another noticeable feature from Figure 5.1 is the zigzag pattern in the monthly admissions, which pulled down admission numbers during December and January. This pattern can be attributed to the fact that during the holiday season, members may not opt for elective surgeries, because as private patients they have the privilege to choose times for elective surgeries. The pattern may also be attributed to the fact that the hospitals may be understaffed for providing non-emergency services during these periods.

![Number of admission per month in whole dataset](image)

**Figure 5.1: Monthly admission statistics for entire dataset**

After understanding these basic properties of the overall population, our next focus was to organise the dataset according to our predefined structure. As described in Chapter 3, like any other raw healthcare dataset, the original PostgreSQL format had some integrity issues, and data were segmented over multiple tables. The dataset also lacked strict
primary key–foreign key relationships across tables, and database indexes were not properly formed, which resulted in slow read performance. Further, the framework requires sending frequent read and write requests to the native database. As a result, we needed to transfer the required data into a Microsoft SQL Server database and organise it according to the research’s specific format. We will refer to the data that we fetched and transferred from the original database to the new SQL server as the research database.

To do most of the database filtering, integrity checking and framework implementation, we developed a custom software. The software\(^3\) was built upon the .NET framework in C# programming language, and thus is capable of leveraging the built-in and extensively supported interface with Microsoft SQL Server, in which we hosted the research database. The first task of the software was to check for data integrity in the original database, to filter out the patients who either did not have any admissions or did not have any diagnosis information during their admissions. Next, the software identified the diabetic patients by looking for the diabetes-related ICD codes in the diagnosis information in the admission records of the valid patients. The ICD codes for diabetes are E10 to E14, so anyone whose diagnosis information begins with at least one of these codes was flagged as diabetic in the original dataset. We found a total 18,353 such patients. As we needed an equal number of diabetic and non-diabetic patients, and out of these estimated 18,000 patients only a smaller sample would be used for the framework, we did not need to include all the available non-diabetic patients. Therefore, we randomly sampled around 14,000 non-diabetic patients. The flagged cohorts of diabetic and non-diabetic patients were then copied from the original database to the research database using the software. The research database thus contained patients’ basic demographic information and complete admission history along with diagnosis codes. Table 5.2 shows the basic statistics of the research dataset.

\(^3\) Please refer to the Appendix F and J for further details of the software and screenshot of the interface.
Table 5.2: Basic statistics for research dataset

<table>
<thead>
<tr>
<th>Total records</th>
<th>Overall</th>
<th>Diabetic (all forms)</th>
<th>Non-diabetic</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Overall</strong></td>
<td>32,820</td>
<td>18,353</td>
<td>14,467</td>
</tr>
<tr>
<td><strong>Number of patients</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Male</strong></td>
<td>16,755 (51%)</td>
<td>10,368 (56.5%)</td>
<td>6,387 (44.15%)</td>
</tr>
<tr>
<td><strong>Female</strong></td>
<td>16,065 (49%)</td>
<td>7,985 (43.5%)</td>
<td>8,080 (55.85%)</td>
</tr>
<tr>
<td><strong>Number of admissions</strong></td>
<td>32,1249</td>
<td>48,790</td>
<td>272,459</td>
</tr>
<tr>
<td><strong>Average admissions per patient</strong></td>
<td>9.79</td>
<td>2.658</td>
<td>18.833</td>
</tr>
</tbody>
</table>

From Table 5.2, the first noticeable difference is the male-to-female ratio in the non-diabetic patients. Recall that in the original dataset, the male-to-female ratio (see Table 5.1) was about 45.3% to 54.7%. A similar ratio was found among the sampled non-diabetic patients in the research dataset. However, for the diabetic patients, the opposite is observed. The male population (56.5%) is larger than the female population (43.5%). If we look at known statistics on the prevalence of diabetes between genders, according to the Australian government (AIHW, 2016), men have a higher prevalence (60%) than women (40%). These statistics are for 2011–12 and include type 1, type 2 and type unknown diabetes, and exclude gestational diabetes, which is analogous to our diabetic population in the research dataset. Other research (Logue, et al., 2011) has suggested that men are diagnosed with type 2 diabetes at lower BMIs than women, which may contribute to the higher prevalence among men than women. As type 2 diabetes is the dominant type in our diabetic population, it is logical that this higher prevalence among males carries over to our dataset. Importantly, this difference also suggests that ‘sex’ may be an indicator for predicting risk. However, we need to do further analysis to determine whether or not this difference is really significant for our population. We did this analysis later on when testing the influence of different predictors (e.g., age, sex) in predicting risk for diabetes.
Having organised the research dataset, we now needed to filter out irrelevant or cumbersome records and select the cohorts for the analysis. Two distinct cohorts were required, one of diabetic patients and another of non-diabetic patients. We kept the cohort size the same for both cohorts so that we would not need to normalise the baseline networks by population size at a later stage of the analysis. Now, to select the cohort of diabetic patients from the research dataset, we had a total of 18,353 patients from which to choose. These patients have a different distribution of diabetes-related ICD codes (i.e., E10 to E14) on which they were initially identified. Table 5.3 shows the prevalence of the different types of diabetes according to the ICD codes. The type 2 diabetic condition is more prevalent than type 1 in an 8:1 ratio, matching the Australian statistics for 2011–12 (AIHW, 2016), which found that around 84.9% of the estimated diabetic population has type 2 diabetes and 11.9% has type 1 diabetes. The table also shows a slightly higher prevalence of type 1 and 2 diabetes among males, especially type 2. Other types of the condition have very little significance of representation compared to type 1 and type 2. Malnutrition-related diabetes is a rare condition, and the Australian population, especially those in the dataset (i.e., privately insured), do not tend to be malnourished. Therefore, as expected, malnutrition-related diabetes is absent.

Table 5.3 (next page) shows only the absolute numbers for the different types of diabetes, not the comorbidities between them, as one can have multiple types of diabetes. We have explored these comorbidities and report them in Figure 5.2 (next page). The figure shows people having only type 2 diabetes (15,918 in total) as the dominating group, followed by individuals having only type 1 diabetes (1,357 in total). A smaller number of people (669 in total) have both type 1 and type 2 diabetes. Apart from this group, other combinations of different diabetic conditions are not significantly represented in the dataset, as suggested by Figure 5.2.
Table 5.3: Prevalence of different types of diabetes

<table>
<thead>
<tr>
<th>ICD code</th>
<th>Description</th>
<th>Prevalence</th>
<th>Overall</th>
<th>Male</th>
<th>Female</th>
</tr>
</thead>
<tbody>
<tr>
<td>E10.*</td>
<td>Insulin-dependent (type 1) diabetes mellitus</td>
<td>2,085</td>
<td>1,110</td>
<td>975</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>(53.24%)</td>
<td>(46.76%)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>E11.*</td>
<td>Non-insulin-dependent (type 2) diabetes mellitus</td>
<td>16,812</td>
<td>9,584</td>
<td>7,228</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>(57%)</td>
<td>(43%)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>E12.*</td>
<td>Malnutrition-related diabetes mellitus</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>E13.*</td>
<td>Other specified diabetes mellitus</td>
<td>135</td>
<td>62</td>
<td>73</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>(45.9%)</td>
<td>(54.1%)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>E14.*</td>
<td>Unspecified diabetes mellitus</td>
<td>279</td>
<td>183</td>
<td>96</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>(65.6%)</td>
<td>(34.4%)</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Figure 5.2: Comorbidity pattern within different forms of diabetes

For the diabetic patient cohort, we only considered type 2 diabetes (T2D) as the chronic disease, and the subsequent framework, therefore, predicts risk for T2D only. The reason for this is that T2D has known risk factors and is potentially preventable or manageable at early stages of detection, therefore making it suitable for analysis and risk prediction.
modelling. In addition, T2D is the most prevalent of the different types of diabetes, as shown in Figure 5.2 and Table 5.2. In contrast, type 1 diabetes is a type of autoimmune disease whose causes are not definitely known; therefore, we avoided including it in the framework. Other types of diabetes (e.g., under E13 and E14 code groups) were also excluded, as their prevalences are quite low and their exact specifications are unknown. Thus, we now had 16,812 T2D patients from which to choose the diabetic cohort and 14,467 non-diabetic patients from which to choose the non-diabetic cohort. The filtering process was performed based on a set of rules that are described in the next section.

5.2 Criteria for Cohort Selection

In this section, we describe the rules or criteria by which we chose the diabetic (type 2) and non-diabetic cohorts from the research dataset. These rules were set after we explored the fundamental properties of the patients in the research dataset, including the number of admissions per patient, time between first and last admissions, and ICD or diagnosis codes present in the admission histories. During this assessment, we tried to determine the quality of the research dataset. It was found that not all patients had sufficient information to be included in the cohorts. Therefore, based on these assessments, we determined the rules that were sequentially applied to obtain the cohorts. We briefly list these rules in the following subsections. Note that from this point on, when we mention the ‘diabetic cohort’, we always refer to the cohort of type 2 diabetes patients.

5.2.1 Criteria for selecting the diabetic cohort

1. The patient must be diagnosed with type 2 diabetes: that is, he must have the ICD code E11 in the diagnosis list for any admission. The admission date when the ICD code E11 first appeared is considered as the ‘date of first diagnosis’ in subsequent analysis.

2. The patient must have at least two or more admissions before the ‘date of first diagnosis’. The reason for this is that without two consecutive admissions, we
cannot construct transitions between comorbidities across admissions, making it impossible to generate an individual disease network. Also, the admissions that are counted must have at least one valid ICD code. Due to the coding practices of hospitals or for some same-day admissions, ICD codes were not found for a small number of admissions. Without any ICD codes, there is no point in including an admission, and therefore such admissions were not considered.

3. There should be at least two months (60 days) between the first admission (containing ICD codes) and the ‘date of first diagnosis’. This gap ensures that the individual disease network represents a meaningful time frame for observing disease development. On the other end, we did not consider any maximum threshold for the timeline. The research database itself represents around five and half years’ effective duration (September 2009 to March 2015) between which the admissions occurred. This by default acts as the upper threshold for time duration.

4. The total number of admissions per year should be less than or equal to 30. Anyone having a greater rate than that was excluded from the cohort. Both number of admissions and time are counted between the first valid admission and the ‘date of first diagnosis’. The reason behind this rule is that some patients may require frequent admissions for ongoing treatments such as chemotherapy or kidney dialysis. These recurrent admissions are not accompanied by unique diagnoses or illnesses, but rather represent a course of treatment for a single underlying cause. Including these frequent admissions would disproportionately increase related comorbidities such as cancer or kidney diseases and create bias. Therefore, we assessed patients having chemotherapy or dialysis-related admissions by looking at the number of admissions they had and the time range over which these admissions took place. It was observed that a threshold rate of 30 admissions per year would mostly eliminate this group of patients; therefore, this rate was chosen.
5.2.2 Criteria for selecting the non-diabetic cohort

1. The patient must not be diagnosed with diabetes. For clarity, this includes not only type 2, but all types of diabetes. It means that diabetes-related ICD codes E10–E14 should never be present in the diagnosis list of any admission for non-diabetic cohort patients.

2. As in the diabetic cohort, patients in the non-diabetic cohort must have at least two admissions with valid ICD codes. As non-diabetic patients do not have a ‘date of first diagnosis’, their entire admission history is considered to obtain the total number of effective admissions.

3. As in the diabetic cohort, patients in non-diabetic cohort also must have a minimum amount of time to be considered for the cohort. This is set as two months or 60 days between their first and last admissions. This duration is calculated from the ‘date of admission’ of the first admission until the ‘date of separation’ of the last admission.

4. As in the diabetic cohort, we considered the effect of frequent admissions related to ongoing services such as chemotherapy or dialysis and removed those patients from the cohort. However, for the non-diabetic patients, we chose the threshold slightly differently. Rather than choosing a threshold rate of admissions, we chose a ratio of admissions having ICD codes of interest. These codes, as listed in Table 5.4, indicate healthcare services that require a high volume of admissions for a course of treatment. Note that, in addition to dialysis and chemotherapy, we included rehabilitation care as an ICD indicating frequent admissions, because we found that often a significant number of recurrent admissions would occur for rehabilitation care that may include physiotherapy or occupational therapy for the same course of treatment. Now, to determine the threshold ratio of these types of admissions, we first counted the number of admissions having those ICD codes of interest within the full admission history, then we divided that number by the total number of admissions to obtain the ratio. We observed that, for patients requiring frequent admissions due to ongoing treatments such as chemotherapy, this was
generally greater than 75%. Therefore, for the non-diabetic cohort, we only included patients with frequent admission ratios less than or equal to 75%.

Table 5.4: List of ICD codes to determine frequent admissions

<table>
<thead>
<tr>
<th>ICD codes</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Z49.1</td>
<td>Haemodialysis</td>
</tr>
<tr>
<td>Z51.1</td>
<td>Pharmacotherapy session for neoplasm (chemotherapy)</td>
</tr>
<tr>
<td>Z50.9</td>
<td>Care involving use of rehabilitation procedure, unspecified</td>
</tr>
</tbody>
</table>

We chose the eligible patients for the diabetic and non-diabetic cohorts by sequentially applying the above-mentioned criteria. Then, among the eligible patients, we randomly sampled 2,300 diabetic patients as the diabetic cohort and an equal number of non-diabetic patients as the non-diabetic cohort. A multiple of ten was chosen because we wanted to apply k-fold validation with $k = 10$ on the cohorts, requiring that the cohorts be divisible into ten equal-sized partitions. At this point, all cohorts had sufficient information for each patient and were ready to be used in the analysis. We now discuss some key characteristics of the final cohorts.

5.3 Characteristics of the Diabetic and Non-diabetic Cohorts

Figure 5.3 shows the age distribution of the diabetic and non-diabetic cohort patients. For the diabetic cohort, age was calculated up to the date of first diagnosis. For the non-diabetic cohort, age was calculated up to the mid-point in time of the data duration; that is, if a non-diabetic patient’s dates of first and last admission are 1 January 2010 and 31 December 2011, then the mid-date of 31 December 2010 is used to calculate that patient’s age.
Figure 5.3: Age distribution of the diabetic and non-diabetic cohorts

Figure 5.3 shows that the age distribution of the diabetic cohort is slightly amassed to the right compared to the non-diabetic cohort. It also shows that most diabetic patients experience onset of disease at 45+ years, peaking around 60 years. This is analogous with the overall Australian statistics (AIHW, 2016), which show that the prevalence rate of type 2 diabetes is 5% for 45–54 year-olds and 16% for 55–64 year-olds, almost three times higher. For the non-diabetic cohort, frequency in age begins to increase somewhat earlier, at around age 35, and peaks around 60–70 years. The older age distribution for the diabetic cohort suggests that age can be an indicator for predicting risk. However, some part of the distribution overlaps across both cohorts, which also suggests that patients within this range may incorrectly be predicted if we consider age as the only risk factor. Later on, we discuss the tests performed to identify whether or not the difference in age distribution is significant, and attempt to determine the effectiveness of age as a risk predictor.
Figure 5.4: Time between first and last considered admission for diabetic and non-diabetic cohorts

Figure 5.4 gives an idea of the distribution across the two cohorts of the effective time period for the analysis. For the diabetic cohort, it shows the time from the first admission to the date of first diagnosis. This time period can be treated as lookback period during which we have the data to understand what disease codes are appearing in those diabetic patients leading to their diagnoses. For diabetic cohort, the figure shows a slightly fluctuating trend over different lookback periods. The average lookback period for the diabetic cohort is about 2.5 years (29.5 months). For the non-diabetic cohort, a similar trend is observed.
5.4 Code Range Selection

Having identified our diabetic and non-diabetic cohorts, the next step was to generate the two baseline networks from each of them. Before doing so, we needed to find the appropriate scope of the ICD code ranges for the analysis. We found around 7,000 different ICD codes in the diagnosis lists accompanying admission records. As discussed in Chapter 4, many of these ICD codes do not indicate particular diagnoses or diseases; rather, they specify other attributes, such as the place of the incident. Also, some diagnoses, like fever or vomiting, represented in forms of the ICD code, are not necessarily relevant to predicting diabetes risk. Therefore, we have two options—one is to select a group of ICD codes that are relevant, and the other is to consider all ICD codes up to the first three characters. We discuss these options in the subsections below.

5.4.1 Selective comorbidities based on the Elixhauser index

We used selected comorbidities as the main method of grouping ICD code ranges in the framework. In this method, we used a translation table that maps each comorbidity with its corresponding ICD codes. If any ICD codes are matched with entries in the diagnosis list of a patient’s admission history, the corresponding comorbidity is flagged as present for that admission. Therefore, the nodes in the baseline networks are not the individual ICDs, but rather comorbidities that indicate that the patient has at least one or more ICD codes related to that comorbidity. We used the Elixhauser index as the basis for the comorbidities. The original index had 30 comorbidities, later modified to 31 comorbidities (Garland, et al., 2012). Among these 31 comorbidities, two were related to diabetes. Because our goal is to predict risk of diabetes, we excluded these two comorbidities. Table 5.5 shows the 29 comorbidity categories. Detailed lists of the corresponding ICD codes that represent these comorbidities are given in the Appendix G.
Table 5.5: List of Elixhauser comorbidities used in the framework

<table>
<thead>
<tr>
<th>S/L</th>
<th>Comorbidity</th>
<th>S/L</th>
<th>Comorbidity</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Congestive heart failure</td>
<td>16</td>
<td>Lymphoma</td>
</tr>
<tr>
<td>2</td>
<td>Cardiac arrhythmias</td>
<td>17</td>
<td>Metastatic cancer</td>
</tr>
<tr>
<td>3</td>
<td>Valvular disease</td>
<td>18</td>
<td>Solid tumour without metastasis</td>
</tr>
<tr>
<td>4</td>
<td>Pulmonary circulation disorders</td>
<td>19</td>
<td>Rheumatoid arthritis/collagen vascular diseases</td>
</tr>
<tr>
<td>5</td>
<td>Peripheral vascular disorders</td>
<td>20</td>
<td>Coagulopathy</td>
</tr>
<tr>
<td>6</td>
<td>Hypertension, uncomplicated</td>
<td>21</td>
<td>Obesity</td>
</tr>
<tr>
<td>7</td>
<td>Hypertension, complicated</td>
<td>22</td>
<td>Weight loss</td>
</tr>
<tr>
<td>8</td>
<td>Paralysis</td>
<td>23</td>
<td>Fluid and electrolyte disorders</td>
</tr>
<tr>
<td>9</td>
<td>Other neurological disorders</td>
<td>24</td>
<td>Blood loss anaemia</td>
</tr>
<tr>
<td>10</td>
<td>Chronic pulmonary disease</td>
<td>25</td>
<td>Deficiency anaemia</td>
</tr>
<tr>
<td>11</td>
<td>Hypothyroidism</td>
<td>26</td>
<td>Alcohol abuse</td>
</tr>
<tr>
<td>12</td>
<td>Renal failure</td>
<td>27</td>
<td>Drug abuse</td>
</tr>
<tr>
<td>13</td>
<td>Liver disease</td>
<td>28</td>
<td>Psychoses</td>
</tr>
<tr>
<td>14</td>
<td>Peptic ulcer disease excluding bleeding</td>
<td>29</td>
<td>Depression</td>
</tr>
<tr>
<td>15</td>
<td>AIDS/HIV</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

We used the ICD codes for the above 29 comorbidities from the listing provided by Quan et al. (2005). The original task was to translate the earlier version of the ICD-9-CM code for comorbidities and define a newer ICD-10 coding algorithm. Because the ICD standard used in our dataset is ICD-10-AM, we manually examined the ICD codes in the original listing against the corresponding ICD-10-AM codes to check whether any definitions had been updated. However, all the ICD codes in the original listing were the same as those in the ICD-10-AM code dictionary. Therefore, we did not need to modify the original ICD listings for the 29 comorbidities.

Along with the Elixhauser comorbidities, we studied the literature to find what specific comorbidities and conditions are often prevalent in diabetic patients. The reason for doing
so is that the Elixhauser index was designed to predict in-hospital mortality. The chronic comorbidities of the index are obviously relevant in predicting diabetes risk, but they may exclude conditions that are related to diabetes but not to the task of predicting in-hospital mortality. Upon critical examination, we added six more comorbidities and health conditions, bringing the overall number of comorbidities to 35. The newly added comorbidities are listed in Table 5.6 with their corresponding ICD-10-AM codes alongside.

Table 5.6: Comorbidities and health conditions added to Elixhauser index

<table>
<thead>
<tr>
<th>S/L (continuing from previous table)</th>
<th>Comorbidity or health condition</th>
<th>ICD-10-AM</th>
</tr>
</thead>
<tbody>
<tr>
<td>30</td>
<td>Cataract</td>
<td>H25?, H26?</td>
</tr>
<tr>
<td>31</td>
<td>Anaemia, unspecified</td>
<td>D64.9</td>
</tr>
<tr>
<td>32</td>
<td>History of long-term medication, insulin</td>
<td>Z92.22</td>
</tr>
<tr>
<td>33</td>
<td>Macular degeneration</td>
<td>H35.3</td>
</tr>
<tr>
<td>34</td>
<td>Presence of coronary angioplasty implant and grafts</td>
<td>Z95.5</td>
</tr>
<tr>
<td>35</td>
<td>Presence of aortocoronary bypass graft</td>
<td>Z95.1</td>
</tr>
</tbody>
</table>

‘?’ indicates wildcard mask: that is, ‘H25?’ means any ICD codes starting with ‘H25’ (e.g., H25.0, H25.1).

In the above table, ‘Presence of coronary angioplasty implant and grafts’ and ‘Presence of aortocoronary bypass graft’ do not primarily indicate any comorbidity. Rather, they are health conditions that indirectly indicate underlying comorbidities for the condition that occurred. For example, these two conditions indicate that the patient had comorbidities like angina, ischemia or coronary heart disease that can lead to heart attack. These comorbidities are critical in predicting diabetes risk, and even though they are not explicitly coded, we can indirectly identify them by looking for the presence of these health conditions. Also, we did not merge these two health conditions with the heart-related comorbidities already present in the Elixhauser index because these conditions do not occur at the same time, and we wanted to keep them logically separated.
As shown in Table 5.6, we also included long-term history of insulin treatment as a comorbidity condition. This was deemed necessary because we found a large number of patients in our diabetic cohort with this condition, and it has the potential to be an indicator. It should be noted that when someone takes insulin injections, it means that person is already diabetic. However, for our diabetic cohort, the effective timeframe considered was that before the patients were diagnosed with type 2 diabetes. The question then arises: why do these patients have a history of insulin medication before being diagnosed with type 2 diabetes? The most probable answer is that these patients had type 1 or another type of diabetes before being diagnosed with type 2, and needed to take insulin for their pre-existing, non-type-2 diabetes. This answer is particularly likely for type 1 diabetic patients, because type 1 is an autoimmune disease in which the pancreatic beta cells cannot naturally produce insulin, and therefore, patients always need to take insulin externally. Chances are that while treating with insulin for the other type of diabetes, these patients also developed type 2 diabetes, and hence it is justifiable to include insulin usage history in the comorbidity list. There is one limitation of this inclusion, which is related to coding quality. There may be chances that the patient was taking insulin solely for treating type 2 diabetes—because a portion of type 2 diabetes patients do really need to take insulin—but that those patients, despite having type 2 diabetes, were not registered with the corresponding ICD code (e.g., E11) to indicate that they are type 2 diabetic, while their insulin usage was nevertheless registered. It is possible that this occurred because of the policy implementation issues that were discussed earlier (see Section 3.3.2). Nevertheless, in both cases insulin usage can function as a predictor for type 2 diabetes risk prediction, and hence we included it.

5.4.2 ICD code collapse

As an alternate to the selected comorbidities method, we used the ICD codes to construct individual disease networks and the baseline network. This method is not used in the main framework, but rather we used it later to provide a comparison measure of performance against the performance results from the main method. The ICD codes were collapsed to
three characters indicating the core diagnosis. The remaining suffixes containing supplementary information about, for instance, the disease’s aetiology, anatomical site or severity were removed. In the diabetic and non-diabetic cohorts, we found 1,600 unique ICD codes when suffixes are ignored, which represents an 81.4% reduction from the total of 8,634 unique ICD codes when suffixes are included. Thus, this method can still capture individual diseases and understand their interrelations, but significantly reduces the volume of nodes in the overall baseline network. Unlike the selected comorbidity list method, the method of creating the baseline network with three-character ICD codes does not demand a manual selection of code groups. This creates a generic advantage over the selected comorbidity list, as the framework utilising the ICD code collapse method can be implemented without a clinical study of the associated comorbidities. On the other hand, because we are not filtering out or merging ICDs into groups, the code range is not fine-tuned, and therefore may introduce inaccuracies. As the number of nodes is quite large when considering individual ICD codes up to three characters, visualisation of the baseline network can be complex compared to the selective comorbidities method. Also regarding execution time, selective comorbidity is faster to implement than using collapsed ICD codes. The exact influence of choosing selective comorbidity or three-character ICD codes is measured later on in the performance and validation section (see Section 5.9).

5.4.3 Selection for socio-demographic factors

In addition to the chosen diagnosis-related ICD codes, we chose three additional discrete and individual-level factors to include in the predictive model. These are (a) age, (b) sex and (c) behavioural risk factors.

Age was the first individual-level risk factor we considered. As described in Section 5.3, for the diabetic cohort, age was calculated using the date of first diagnosis. For the non-diabetic cohort, age was calculated using the mid-point in data duration (i.e., the date halfway between the date of the first admission and the separation date of the last admission). We note that age was only specified in years, because month and date of birth
were obfuscated for data privacy; nevertheless, age in years is sufficiently precise for our analysis. After finding the ages, we normalised the age scores within a range of 0 to 1. We identified the maximum age as near to 110, and therefore divided all the individual ages by 110 to obtain the age score to ensure that all age scores were within the range of 0 to 1.

The second individual-level risk factor was sex. This risk factor is essentially a categorical dichotomous variable, as it can have only two discrete values: male or female; however, it can be interpreted in different ways depending on the methods. In the validation part of the analysis, where binary logistic regression is used, the method automatically converts these categorical values to dummy variables, assigning 1 for male and 2 for female. For the binary classification method, variables are considered categorical. For the default method used in the framework, the parameter estimation model, we converted the sex categories into a dummy variable with -1 for male and +1 for female. The parameter estimation model then finds a suitable weighting factor between -1 to 1 to indicate the influence of sex in predicting T2D risk. The interpretation of the optimal value is as follows: when it is less than 0 and nearer to -1, the male risk score of -1 is multiplied by the negative factor, thus giving an overall positive value for male risk and negative value of female. This indicates higher risk association for males and lower risk association for females. Similarly, if the weighting factor is greater than 0 and nearer to 1, the female group has the higher risk association. If the sex does not have much influence in predicting risk, the weighting factor should give a value close or equal to 0, thus nullifying the overall sex score.

The third risk factor was the behavioural risk factor. Like sex, behaviours are also considered as a categorical dichotomous risk factor, having two possible values: yes or no. However, there is a small difference in the overall combinations of sex and behavioural risk factors in predicting risk. The sex-based risk factor has three possible outcomes: the male group has a higher risk, the female group has a higher risk, or sex does not have any influence. For the behavioural risk factor, there are only two possibilities: either the
behaviour has an influence on overall risk or it does not. Therefore, in our parameter estimation model, we assigned 1 if the patient has behavioural risk factors present and 0 otherwise. Binary logistic regression internally uses 1 and 2 as the dummy variable to indicate absence or presence of the risk factor, while binary tree classification uses the categorical variable as it is. In our framework, we examined smoking or tobacco use as the behavioural risk factor. Two ICD codes were checked to determine the presence of the risk factors. The first one was Z86.43, which indicates ‘Personal history of tobacco use disorder’, and the second was Z72.0, which indicates ‘Tobacco use (current)’. If either of these two ICD codes are present in any of the admissions within the effective lookback time for the patient, that patient’s behavioural risk factor is flagged as ‘yes’. Table 5.7 provides a summary of individual risk factors.

### Table 5.7: Individual risk factors related to demographics and behaviour

<table>
<thead>
<tr>
<th>Risk factor</th>
<th>Type of variable</th>
<th>Possible values</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age</td>
<td>Continuous</td>
<td>0 to 1, normalised by dividing age (in years) by 110 (maximum age).</td>
</tr>
<tr>
<td>Sex</td>
<td>Categorical</td>
<td>In parameter estimation model, -1 for male, 1 for female.</td>
</tr>
<tr>
<td></td>
<td>dichotomous</td>
<td>In binary logistic regression 1 for female, 2 for male.</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Binary tree classification use categorical values as they are.</td>
</tr>
<tr>
<td>Behaviour</td>
<td>Categorical</td>
<td>In parameter estimation model, 1 for presence of factor, 0 otherwise.</td>
</tr>
<tr>
<td></td>
<td>dichotomous</td>
<td>In binary logistic regression 1 for absence, 2 for presence.</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Binary tree classification use categorical values as they are.</td>
</tr>
</tbody>
</table>
5.5 Comparison Across Three Baseline Networks

At this stage of the analysis, we generated the positive baseline network from the diabetic patient cohort and the negative baseline network from the non-diabetic patient cohort. As discussed earlier, we used selective comorbidities that grouped related ICD codes into one comorbidity category. Therefore, the nodes in these two baseline networks represent comorbidities and the edges represent transitions between comorbidities in subsequent admissions. Overall, these two baseline networks represent the health trajectories of the diabetic and non-diabetic patients. From these two networks, we calculated the difference by attribute adjustment and generated the final and summary baseline network\(^4\) that represents the unique properties of diabetic patients. We then calculated several network and corresponding node-level measures to understand the features of these three baseline networks. In this section, we discuss the findings from these networks. Figure 5.5 shows a block diagram of the three networks leading to the network and node level comparison.

---

\(^4\) Henceforth, we will refer this as final baseline network. Later on, when we discuss the prediction, we will simply refer this as baseline network, as this is the only network present at that stage. Therefore, whenever we simply mention Baseline Network in this thesis it will always refer to final baseline network.
First, we present the network-level measures of three baseline networks in Table 5.8. It is important to note that we used k-fold \((k = 10)\) validation, as discussed in Chapter 4 (see Section 4.11.1). Therefore, we divided each cohort into 10 equal-sized partitions, then generated the baseline networks and subsequent risk prediction and validation components 10 times, each time using a different combination of partitions. This means that the baseline networks had 10 different versions as generated from the different partitions. It is not practical to report each of them individually; besides, the statistics were found to be quite similar across each iteration (see Section 5.9 for comparison). Therefore, we report the results of the baseline networks as obtained from the first fold (i.e., fold-0). In this fold, the first 6 partitions out of 10 were used to generate baseline networks that include 1,380 (i.e., \(230 \times 6\)) diabetic patients from the diabetic cohort and an equal number of non-diabetic patients from the other cohort.

Table 5.8: Network properties of three Baseline Networks (BN)

<table>
<thead>
<tr>
<th>Criteria</th>
<th>Positive BN</th>
<th>Negative BN</th>
<th>Final (summary) BN</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of nodes</td>
<td>34</td>
<td>33</td>
<td>34</td>
</tr>
<tr>
<td>Number of edges</td>
<td>658</td>
<td>367</td>
<td>658</td>
</tr>
<tr>
<td>Modularity</td>
<td>0.197</td>
<td>0.345</td>
<td>.045 to .055</td>
</tr>
<tr>
<td>Number of communities</td>
<td>3</td>
<td>6</td>
<td>3</td>
</tr>
<tr>
<td>Network diameter</td>
<td>3</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>Average path length</td>
<td>1.453</td>
<td>1.654</td>
<td>1.453</td>
</tr>
<tr>
<td>Graph density</td>
<td>0.586</td>
<td>0.348</td>
<td>0.586</td>
</tr>
<tr>
<td>Average Clustering Co-efficient</td>
<td>0.72</td>
<td>0.579</td>
<td>0.72</td>
</tr>
</tbody>
</table>

Table 5.8 shows that all three baseline networks show an almost equal number of nodes. This is not surprising, because we had already selected the groups of ICD codes that would be translated into single comorbidities. This increased the chance that there would be at least one ICD code from each of the comorbidities from each of the cohorts, and therefore that all three baseline networks would have the comorbidity. Among the 35 selective
comorbidities, HIV/AIDS is absent from the whole dataset, which is also normal. The negative baseline network also does not have the node indicating ‘long-term insulin use’, and thus has one less node than the other two. This is fitting with our previous discussion on including this health condition in the list. Insulin intake is a good predictor of type 2 diabetes, as it means the patient has some type of diabetes. For our present case, it means the patient has type 1 or any other type of diabetes excluding type 2. (type 1 is the most probable case as it is more prevalent). For these types of diabetes, they are taking insulin indicating the patients already have metabolic problems. These conditions eventually lead to type 2 diabetes, thus making insulin intake a good predictor. Even if the patient actually was taking insulin solely for the purpose of managing a case of type 2 diabetes not recorded in the admission data until after some period of insulin intake, the positive and negative baseline networks correctly reflect this—no non-diabetic patient takes insulin, and those who take insulin to manage other types of diabetes eventually develop type 2 diabetes, therefore making it a good predictor.

The edge count in the positive baseline network is almost double that of the negative baseline network. This indicates that the positive baseline network is more dense or saturated: that is, diabetic patients have more transitions between comorbidities in subsequent admissions. This also suggests that diabetic patients have relatively more admissions and/or more exclusive comorbidities in subsequent admissions, thus increasing the edge count. The high graph density value for the positive baseline network also strengthens this fact. Graph density is the measure of edge saturation: that is, how many edges are present compared to the total possible number of edges. For the positive baseline network, the graph density is 58.6%, indicating that almost 6 out of every 10 possible transitions are present, which is nearly double the rate of the negative baseline network (i.e., 34.8%, see Table 5.8 above). This high number of transitions in the positive baseline network indicates that diabetic patients represent a higher admissions burden and complex progression pathways over subsequent admissions.
We see an opposite picture for the modularity property: the negative baseline network has almost double the modularity of the positive one. Modularity is the measure of clustering, or the community formation tendency among the nodes. A cluster or community is a group of nodes that are closely interconnected compared to the others. It is expressed as a value between 0 and 1 inclusive, where 1 indicates the highest modularity and 0 indicates no modularity. To obtain a higher modularity score, the network would normally have some groups of nodes that have a high number of edges among themselves and a lower number of edges with nodes of different groups. As we have observed, diabetic patients have a large number of (58.6% of all possible) transitions between comorbidities; therefore, there is less chance that some nodes will have dense transitions between themselves and sparse transitions with others. In contrast, there are fewer transitions for the non-diabetic patients; therefore, they have a higher chance of group or community formation, and hence have a higher (almost double) modularity than the positive baseline network. Nevertheless, from a network standpoint, both modularity scores are still quite small, probably because we pre-fixed the maximum number of nodes (i.e., comorbidities) to 35. We may encounter higher modularity where the node size is larger. Apart from these two baseline networks, we now see the measure of the third network i.e., the final baseline network (to recall this network, please see Figure 5.5 at the beginning of this section). This final baseline network has a very small modularity as evident from the Table 5.8. This is because we generated this network via attribute adjustment. The edge frequency (i.e., edge attribute), which indicates frequency of transitions in both positive and baseline networks and was actually a large number in many cases, was normalised between 0 and 1. As the modularity algorithm takes edge frequency into account, the final baseline network received a lower modularity score. In reality, the final summary baseline network is not structurally different from the positive baseline network, as the summary baseline network is essentially a set union of nodes and edges of positive and negative baseline networks. For this research, we can observe that all of the summary baseline network’s structure was inherited from the positive baseline.
network, with only a portion that has been attribute-adjusted from the negative baseline network.

The number of communities is the outcome of modularity detection. A community is a group of nodes which have relatively more interconnection within themselves compared to the connection with other nodes. As discussed in the previous section, the negative baseline network has higher modularity than the positive one, and we can see that the resultant number of communities (i.e., 6) is also double than that of the positive baseline network (i.e., 3). The final baseline network also generated three communities despite its significantly lower modularity because it has a similar transition structure to the positive baseline network. We should note that modularity does not strictly correlate with the number of communities. Rather, modularity more closely indicates the quality of the communities, that is, how closely knitted the communities are. As we have observed that the baseline networks, especially the final one, have low modularity and a small number of communities, this indicates that the quality or interconnectedness of the communities is not very strong. The overall transition between nodes within the same community is probably not significantly different from the overall transition across inter-community nodes. This gives an indication that the cluster match score, which explicitly examines the similarities with the baseline network’s community structure to predict risk for the test patient, may not be very influential because the baseline network’s community structure is not very well formed. We test this influence in detail in Section 5.7.

The remaining two measures—average path length and average clustering coefficient—do not exhibit important features in the present context. Path length is the distance between two nodes in terms of ‘hops’. For the positive and final summary baseline networks, every node is reachable from other nodes through 1.45 hops on average, and for the negative baseline network, the distance is slightly greater (1.65). This is because the nodes are connected by a large number of edges, thereby creating short-circuits and lowering average path length. It should be noted that relation between edge count and average path length is non-linear and depends on how the edges are attached to the nodes.
Generally, for the real-world networks, there is a certain threshold range, within which the
count quickly lowers the average path length. After that, if more edges
are added, they do not provide many shortcuts to reach other nodes, i.e., do not lower
the average path length much. In our case, we noted before that not all the edges are
present in the three baseline networks and for the negative baseline network it is even
smaller. Nevertheless, the number of edges present is big enough to lower the average
path length to reach very small range (i.e., 1.45 to 1.65) for all three baseline networks.

The next measure - clustering coefficient is the ratio of triangle structures, that is, three
comorbidities having all of the six possible transitions between them. This coefficient is
higher in the positive and summary baseline networks because they have a higher edge
count.

Having examined the network structure of the three baseline structures, we now turn to
discuss the individual-level properties of the baseline networks: the prevalent comorbidity
and transition pattern before and after attribute adjustment is implemented.

5.6 Findings for Attribution Effects

The final baseline network derived from the positive and negative baseline networks
represents unique characteristics of progression. Having considered the attribution effect
across the diabetic and non-diabetic cohorts, the resultant baseline network assigns a
higher weight to comorbidities (i.e., node frequency) and their progression (i.e., edge
frequency) for those that are more prevalent in diabetic patients. In the following
subsections, we look at comorbidities and transitions that are prevalent in the diabetic
and non-diabetic cohorts, and discuss the ones that have attributed more to the final
baseline network. Further, we discuss the community or cluster formation in the baseline
network.
5.6.1 Prevalent comorbidities

Table 5.9 shows the top 10 most prevalent comorbidities in the positive and negative baseline networks. We can easily notice that the top comorbidities listed in the positive baseline network show significantly higher prevalence compared to the negative baseline network. For example, the most prevalent comorbidity in diabetic patients (cardiac arrhythmias) has nearly 7,000 cases, where even the most prevalent comorbidity for the non-diabetic patients has just over 1,200 cases and cardiac arrhythmias did not came in the top list in non-diabetic patients. A similar trend is observed in the other ranks. Because both networks are derived from the same number of patients, this suggests a complex and frequent comorbidity burden for the diabetic patients.

<table>
<thead>
<tr>
<th>S/L</th>
<th>Diabetic patient (+ve baseline network)</th>
<th>Prevalence</th>
<th>Non-diabetic patient (-ve baseline network)</th>
<th>Prevalence</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Cardiac arrhythmias</td>
<td>6,938</td>
<td>Depression</td>
<td>1,284</td>
</tr>
<tr>
<td>2</td>
<td>Solid tumour without metastasis</td>
<td>5,314</td>
<td>Solid tumour without metastasis</td>
<td>1,102</td>
</tr>
<tr>
<td>3</td>
<td>Metastatic cancer</td>
<td>2,955</td>
<td>Metastatic cancer</td>
<td>706</td>
</tr>
<tr>
<td>4</td>
<td>Cataract</td>
<td>2,093</td>
<td>Renal failure</td>
<td>426</td>
</tr>
<tr>
<td>5</td>
<td>Depression</td>
<td>821</td>
<td>Other neurological disorders</td>
<td>411</td>
</tr>
<tr>
<td>6</td>
<td>Hypertension, uncomplicated</td>
<td>712</td>
<td>Weight loss</td>
<td>228</td>
</tr>
<tr>
<td>7</td>
<td>Renal failure</td>
<td>660</td>
<td>Alcohol abuse</td>
<td>187</td>
</tr>
<tr>
<td>8</td>
<td>Anaemia, unspecified</td>
<td>391</td>
<td>Lymphoma</td>
<td>176</td>
</tr>
<tr>
<td>9</td>
<td>Congestive heart failure</td>
<td>387</td>
<td>Drug abuse</td>
<td>159</td>
</tr>
<tr>
<td>10</td>
<td>Alcohol abuse</td>
<td>386</td>
<td>Chronic pulmonary disease</td>
<td>149</td>
</tr>
</tbody>
</table>
From Table 5.9, it is also evident that there are differences between the two networks in terms of the prevalence of the same comorbidities: that is, some comorbidities are more prevalent in one of the two networks. For example, the most prevalent comorbidity in the positive baseline network (cardiac arrhythmias) is absent from the top 10 list of the negative baseline network. Solid tumour without metastasis is almost five times more prevalent in diabetic patients, although both networks have it as the second highest comorbidity. This indicates that these comorbidities are more associated with diabetes. Conversely, some comorbidities are more prevalent in non-diabetic patients (e.g., depression, weight loss), therefore indicating that these are not really associated with determining diabetes risk. Thus, based on the relative differences in individual-level properties—that is, via attribute adjustment of comorbidities and transitions—we assigned the risk factors in the final summary baseline network. The algorithm for attribute adjustment is given in the subsection 4.8. It should be noted that after the attribute adjustment, the score for risk factors are normalised for the final baseline network, i.e., the values are represented in the range between 0 and 1 inclusive. Figure 5.6 shows the top 10 comorbidities or conditions that attributed most for diabetic patients in the final baseline network.
Figure 5.6: Top 10 comorbidities and conditions that attributed most for diabetes patients

The figure shows that the highest weight of 1 is given to ‘cardiac arrhythmias’ and ‘long-term use of insulin’. Cardiac arrhythmias was the most prevalent comorbidity in the positive baseline network, and also had a very small prevalence in the negative baseline network, therefore achieving a high ratio in attribute adjustment. Because this was the maximum ratio, it was given the weight of 1 and all other weights were scaled between 0 and 1 accordingly. As discussed earlier, long-term use of insulin was exclusive to diabetic patients. Therefore, although its prevalence did not make the top 10 list in the positive baseline network, it received the highest prevalence score in the final baseline network. Except for liver disease and cataracts, the other comorbidities received scores of less than 0.2 as prevalence or node attributes. This reflects the fact observed earlier that most
comorbidities in diabetic patients are also present in non-diabetic patients, often with little differentiation.

5.6.2 Prevalent transition patterns

We now look at the transition pattern between the comorbidities in the final baseline network. Table 5.10 shows the most significant transitions, that is, the transitions with the highest edge weight attribute. For this reporting, we included only unique transitions and excluded self-loops, or transitions to the comorbidity itself in subsequent admissions; however, self-loops were kept in the baseline network for risk prediction. The prevalent transitions show many comorbidities are associated with heart-related problems like valvular disease or congestive heart failure. The inter-related comorbidity transition pattern for the diabetic patient is also evident from the table. For example, the most common transition—from renal failure to valvular disease—indicates problems in two different body systems. Renal failure is a kidney disease; for diabetic patients, lack or imbalance of insulin production can lead to high blood sugar levels. This makes the kidneys filter too much blood to get rid of the excess sugar, and the excessive strain can lead to kidney disease (American Diabetes Association, 2013). Factored with old age, late diagnosis and other comorbidities, this can lead to renal failure. On the other hand, valvular disease is a disease associated with the four valves present in the heart (National Heart, 2015). Its risk factors include other heart diseases, obesity and diabetes (Benjamin, et al., 1994). Now, renal failure and valvular disease may appear unrelated, but for diabetic patients this is a common transition, as diabetes is a risk factor for both. As we have observed that the baseline network’s highest edge attribute value is assigned to this progression, this indicates that the transition from renal failure to valvular disease is nearly exclusive to diabetic patients compared to non-diabetic. The other top transitions shown in Table 5.10 indicate similar exclusiveness for diabetic patients, having edge attributes of over .99 (not shown in the table) on a scale of 0 to 1.
Table 5.10: Top 10 most frequent progressions between comorbidities or health conditions in subsequent admissions according to the final baseline network

<table>
<thead>
<tr>
<th>Initial condition</th>
<th>Next condition</th>
</tr>
</thead>
<tbody>
<tr>
<td>Renal failure</td>
<td>Valvular disease</td>
</tr>
<tr>
<td>Deficiency anaemia</td>
<td>Solid tumour without metastasis</td>
</tr>
<tr>
<td>Presence of coronary angioplasty implant and graft</td>
<td>Congestive heart failure</td>
</tr>
<tr>
<td>Solid tumour without metastasis</td>
<td>Other neurological disorders</td>
</tr>
<tr>
<td>Renal failure</td>
<td>Hypertension uncomplicated</td>
</tr>
<tr>
<td>Metastatic cancer</td>
<td>Renal failure</td>
</tr>
<tr>
<td>Hypertension uncomplicated</td>
<td>Presence of coronary angioplasty implant and graft</td>
</tr>
<tr>
<td>Cataract</td>
<td>Congestive heart failure</td>
</tr>
<tr>
<td>Solid tumour without metastasis</td>
<td>Renal failure</td>
</tr>
<tr>
<td>Solid tumour without metastasis</td>
<td>Chronic pulmonary disease</td>
</tr>
</tbody>
</table>

At this point, we provide a visualisation of the final baseline network. As we used selective comorbidities for this network, the number of nodes is restricted to a number small enough that we can show all the node labels. Figure 5.7 shows the final baseline network. We used Gephi (Bastian, et al., 2009), a social network analysis software to generate the figure. In the figure, each node represents a particular comorbidity or health condition and is mentioned on the label. The sizes of the nodes and labels are proportional to their corresponding prevalence. The colours of the nodes and label outlines indicate the cluster they belong to. We discuss cluster membership in the next subsection. We applied a force-directed graph layout to position the nodes in such a way that fewer edge overlaps occur between them. This eventually places nodes with high co-occurrences in closer proximity, thus facilitating the visualisation of cluster formation. The baseline network shows ‘cardiac arrhythmias’ and ‘prediabetes medication’ (insulin usage) dominating the network. The network shows a large number of edges between the nodes, with the colour indicating the cluster they are originating from. Although the edge thickness was set to be
proportional to edge weight, this is not very clear in the visualisation, because a large number of edges have high weights with relatively little difference between them, therefore giving the impression that the edges have an equal thickness in the baseline network.

![Figure 5.7: Baseline network after attribute adjustment](image)

**5.6.3 Clustering membership in baseline network**

We applied cluster or community detection algorithm proposed by Blondel et al. (2008). We have already discussed the algorithm in details in section 4.3.2. The algorithm was built-in the Gephi software which was used for detecting the clusters. We found three clusters in the baseline network. As we discussed earlier in Section 5.5 related to the
network properties of baseline networks, the modularity score is not very high, indicating that the baseline network does not exhibit a strong community structure. This is because diabetic patients have quite complex and extensive disease trajectories, thus making the positive and final baseline network quite saturated. Nevertheless, the algorithm found three distinct communities that have stronger transitions between comorbidities within the group. In the previous Figure 5.7, we saw the membership and approximate structure of the clusters within the baseline network. We now list the individual clustered comorbidities in Table 5.11.

We can see that the clusters are of relatively equal size in terms of comorbidity count. Cluster 1 has the highest number of comorbidities (i.e., 13). The comorbidities within this cluster are mostly related to heart diseases, as we can observe cardiac arrhythmias, hypertension, the presence of bypass grafts and chronic pulmonary disease included in the cluster. We also observe cancer and anaemia-related comorbidities as members of Cluster 1.

Cluster 2 shows the long-term use of insulin and liver disease as the most prevalent comorbidities within the cluster. This membership is significant because of the pathophysiological relations between these two and diabetes. We attempt to describe this briefly. Two vital hormones are secreted by the pancreas that play a significant role in regulating blood glucose level: these are insulin and glucagon. When we have a major meal, blood glucose level rises sharply, and the insulin is released from the pancreas to maintain the balance. Insulin increases glucose uptake in muscle and adipose tissue and promotes glycolysis (i.e., breaking down of glucose) and glycogenesis (i.e., glycogen formation from glucose) in liver and muscle (Szablewski, 2011). Thus insulin helps to lower the glucose level. Conversely, when blood glucose level falls low, glucagon is released that promotes glucose production and elevates glucose concentration to normal. This vital process of balancing insulin and glucagon to maintain blood glucose is called ‘glucose homeostasis’, and the liver plays an important role in it. The liver is the principal site of glucose deposition when insulin is released to lower blood glucose (Sherwin, 1980).
Table 5.11: Cluster membership among comorbidities in the baseline network listed according to the prevalence rate

<table>
<thead>
<tr>
<th>S/L</th>
<th>Cluster 1</th>
<th>Cluster 2</th>
<th>Cluster 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Cardiac arrhythmias</td>
<td>Long-term use of insulin</td>
<td>Cataract</td>
</tr>
<tr>
<td>2</td>
<td>Hypertension, uncomplicated</td>
<td>Liver disease</td>
<td>Presence of coronary angioplasty implant and graft</td>
</tr>
<tr>
<td>3</td>
<td>Presence of aortocoronary bypass graft</td>
<td>Valvular disease</td>
<td>Congestive heart failure</td>
</tr>
<tr>
<td>4</td>
<td>Deficiency anaemia</td>
<td>Macular degeneration</td>
<td>Pulmonary circulation Disorders</td>
</tr>
<tr>
<td>5</td>
<td>Hypertension, complicated</td>
<td>Fluid and electrolyte disorders</td>
<td>Coagulopathy</td>
</tr>
<tr>
<td>6</td>
<td>Solid tumour without metastasis</td>
<td>Obesity</td>
<td>Anaemia unspecified</td>
</tr>
<tr>
<td>7</td>
<td>Metastatic cancer</td>
<td>Alcohol abuse</td>
<td>Peripheral vascular disorders</td>
</tr>
<tr>
<td>8</td>
<td>Blood loss anaemia</td>
<td>Other neurological disorders</td>
<td>Lymphoma</td>
</tr>
<tr>
<td>9</td>
<td>Renal failure</td>
<td>Psychoses</td>
<td>Rheumatoid arthritis/collagen vascular diseases</td>
</tr>
<tr>
<td>10</td>
<td>Paralysis</td>
<td>Depression</td>
<td></td>
</tr>
<tr>
<td>11</td>
<td>Chronic pulmonary disease</td>
<td>Drug abuse</td>
<td></td>
</tr>
<tr>
<td>12</td>
<td>Hypothyroidism</td>
<td>Weight loss</td>
<td></td>
</tr>
<tr>
<td>13</td>
<td>Peptic ulcer disease excluding bleeding</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

When the body is in a fasting state and glucose levels decrease, the liver is solely responsible for the delivery of glucose to the bloodstream. Now, if the body (i.e., muscle, fat and liver) cannot use the insulin effectively, this can lead to insulin resistance, higher blood glucose concentration and eventually type 2 diabetes. In the case of insulin
resistance, beta cells in the pancreas continue trying to lower the glucose level by increasing insulin production, as the normal level of insulin is not working efficiently. At a certain point in time, the beta cells fail to keep up with increased demand for insulin, leading to diabetes (National Institute of Diabetes and Digestive and Kidney Diseases, 2014) and requiring external insulin uptake. On the other end, abnormal glucose homoeostasis can lead to liver diseases because of its central role in the process. Therefore, we can observe the relation between liver disease and insulin uptake, and their prevalence and membership within the same cluster.

Among the other comorbidities in Cluster 2, we can observe the behavioural and related disorders (e.g., depression, psychoses, drug and alcohol abuse). Obesity and weight loss can also be related to depression or drug abuse. Therefore, their inclusion in the same cluster makes sense.

Cluster 3 has the smallest number of comorbidities (i.e., 9). We can notice few heart-related comorbidities and conditions present in this cluster, such as the presence of, for example, coronary angioplasty implants and grafts, congestive heart failure and pulmonary circulation disorders. The other cluster memberships are not of great importance to discuss here.

This concludes the discussion of the findings on the first part of the framework—constructing the baseline network through attribution adjustment. In summary, we used the first 6 of 10 partitions as part of fold-0 to create the baseline network. This gave us a total of 1,380 diabetic and an equal number of non-diabetic patients to form the baseline network. In the next section, we discuss the risk prediction part of the framework.

5.7 Parameter Estimation

In this part, we generate the predictive model and analyse the effect and contribution of parameters in the model. Recall that we have identified a total of six parameters potentially contributing to disease risk: two are demographic (age and sex), one is behavioural (tobacco use status), two are graph similarity-matching (node and pattern
match scores) and the last is social network-based cluster matching. The next step is to generate the predictive model that estimates the various contributions of the parameters in the model.

As discussed in the framework section, we employed three types of predictive modelling: the default parameter estimation model proposed by us, binary logistic regression and binary tree classification. The first two models assume a linear relationship between the parameters, and the influence of the parameters is interpreted via their weighting factors. The third model, the binary tree classification model, is not a linear or strictly mathematical model, but rather is widely used as a data mining method. This classification method determines a sequence of optimum boundary check rules on the parameters to classify them into one of the predicted outcomes. The parameter’s influence is determined as an ‘importance measure’, which is a direct numerical measure. All of the parameter estimation methods used need a training set of patients whose risk outcome is known. As we had divided the cohorts into 10 partitions, we used the 7th, 8th and 9th partitions for training. This gave us 690 (i.e., 230 × 3) diabetic patients from the diabetic cohort and an equal number of non-diabetic patients from the other cohort eligible to be included in the training. Recall that the 1st to 6th partitions (total 2,760 patients from both cohorts) were used to generate the baseline networks. Figure 5.8 shows the part of the framework that performs the analysis related to parameter estimation.
Figure 5.8: Parameter estimation component of the framework

As shown in Figure 5.8, we first compared each patient’s individual disease network in the training network with the baseline network and calculated the graph similarity and cluster match scores. The other demographic and behavioural risk scores do not need a baseline network for comparison, as they are simply individual risk scores. Thus, we obtained six risk scores for each patient whose outcome was known (i.e., we knew whether or not these patients would eventually develop type 2 diabetes). Therefore, in the prediction model, we assume that the ‘actual risk’ is the dependent variable and the six risk factors (i.e., age, sex, behavioural factors, graph node match scores) are independent variables.

Once we had obtained values for the independent variables for each test patient, we tried to explore their relationship with the dependent variables. As the primary step, we generated scatter plots for each of the independent variables with the dependent variables. The scatterplot matrices are shown in Figure 5.9.
In all the scatterplot matrices of the above figure, we have at least one categorical variable (i.e., the dependent variable). Therefore, the scatterplot does not form a continuum. When both of the variables are categorical (e.g., sex vs. outcome and behavioural risk vs. outcome), there are only four spots where the points can be placed. Therefore, we used a jitter function, which randomly shifts the points slightly across the categorical variables’ own axis. We also set small transparencies for the points to distinguish between the overlapping scores. As mentioned, one or both of the scatterplot
variables can be categorical. Therefore, we look for a virtual horizontal division line that can effectively separate the points into the opposite side of the division line. If the points evenly reside on both sides of the division line, it indicates that their relation is not one of directly correlation. For Figure 5.9 (c) and (e), we can see that sex and behavioural risk factors seem to be evenly distributed vertically, and they cannot be separated, as observed from plain inspection. The same applies for the behavioural risk factors, and we can sense that these two risk factors may not be influential in the linear prediction modelling (i.e., in the parameter estimation model and binary logistic regression model). The binary classification may give some importance to these two because it is not strictly a linear model. For the scatterplot of age vs. outcome (Figure 5.9 (a)), we can observe a trend towards higher age risk scores for the diabetic patients. Even so, if we think of a horizontal threshold line, and a large number of age scores belonging to non-diabetic patients fall above the threshold line, these points will give false positive scores if we consider age as the only variable. Next, in Figure 5.9 (b), we can see that the graph node match score vs. outcome shows a stronger trend than the others. It shows that most of the diabetic patients received higher graph node match scores than the non-diabetic patients. And the latter cohort’s node match scores have been gravitated towards the zero line. This gives us the idea that graph node match score should be a good predictor. However, just by looking at the scatterplot, we cannot obtain the exact parameter evaluation and accuracy results. We need to generate the predictive model using the three methods mentioned. In the following subsections, we discuss the contribution of the risk factors as obtained from each of the methods.

5.7.1 Influence of risk factors from parameter estimation model

The parameter estimation model iteratively finds optimal parameter values against each of the independent variables so that the overall scores for the diabetic patients are above a certain threshold and scores for the non-diabetic patients are below that threshold. For this analysis, the independent variables that are categorical were converted to dummy variables. The weighting factor range was selected as -1 to 2 inclusive. It should be noted
that earlier (in section 4.10) we mentioned that weighting factors are normally chosen within -1 to 1 inclusive. And, then the model iterates through possible combinations of weighting factors and finds the optimal weighting factors for each the risk factors. The significance is that if the optimal weighting factor is negative, it indicates corresponding risk factor affects the prediction negatively. In other words, the more the risk factor value, the less chance of having the chronic disease and vice versa. If the optimum weighting factor is 0, that means the risk factor does not have any influence on the prediction. Now, the relation between multiple optimal weighting factors is relative. If a model has two risk factors, then having an optimal weighting factor of 0.9 and 0.3 is same as having optimal weighting factors of 0.3 and 0.1, i.e., as long as the ratio is same between the optimal weighting factors and the threshold for prediction scales accordingly. For simplicity, we normally choose a range of -1 and 1 within which the model looks for optimal values. However, in our experiment, we chose a wider range to start with, i.e., from -1 to 2, and reported as such. Therefore, some of the optimal weighting factors may have become more than 1 while reporting later on. As we mentioned, optimal values are relative, therefore, choosing the range of -1 and 1 during the experiment would not affect the prediction outcome and accuracy in any way. It would just scale down the weighting factors and threshold accordingly. Now, we report the optimum weighting factors for the risk factors in Table 5.12. The result was obtained from the execution of fold-0; other folds gave similar results.
The table shows that the graph node match score made the most important contribution to predicting the risk score, setting a weighting factor of 1.04. This is actually a promising result as graph node match score is one of the fundamental contributions proposed in our thesis. We can recall from the definition of graph node match score (see section 4.9.2, in “Graph node match score” part) that the score increases when the test patient’s comorbidities and their frequencies largely overlap with those in the baseline network. A high importance of this node match score, therefore, indicates that this proposed network-based measure is really influential in prediction. Also, it reinforces that the comorbidities and their frequencies differ significantly between diabetic and non-diabetic patients as captured by our baseline network.

The second most important contribution is attributed to the age score, with a weighting factor of 0.2, which is significantly smaller than the previous one. The reason is probably because age score does not differ as much as node match score does between diabetic and non-diabetic patients. This can be seen from the earlier Figure 5.9 (a). Almost all diabetic patients have age scores higher than 0.4 (equivalent to 44 years, when not normalised). This is analogous with the clinical observation that most diagnoses of type 2 diabetes occur after age 45 (American Diabetes Association, 2014). However, if we look at the age score of non-diabetic patients, we can see that a significant portion of them also have similar high age scores, i.e., greater than 0.4. Therefore, age cannot be a very good

<table>
<thead>
<tr>
<th>Risk factor</th>
<th>Optimal weighting factor</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age</td>
<td>0.201</td>
</tr>
<tr>
<td>Sex</td>
<td>0</td>
</tr>
<tr>
<td>Behaviour</td>
<td>0.0001</td>
</tr>
<tr>
<td>Graph node match</td>
<td>1.04</td>
</tr>
<tr>
<td>Graph pattern match</td>
<td>0.1001</td>
</tr>
<tr>
<td>Cluster match</td>
<td>0.0001</td>
</tr>
</tbody>
</table>
predictor alone, as a significant number of non-diabetic patients are also as old as the diabetic patients. The third highest weighting factor having the value of 0.1 is attributed to the graph pattern match score, which is almost half that of the age risk score. The pattern match score matches the disease transition patterns of test patients with the baseline network. A test patient’s disease progression network is often sparse due to the low number of admission count in the dataset which results in a lower number of transition patterns. Subsequently, the resultant pattern match score might be too much granular for the dataset without capturing significant information. Hence, between diabetic and non-diabetic patients, graph pattern match score did not differ much compared to the other two more significant predictors, i.e., graph node match score and age score.

The other three risk factors did not achieve much significance in terms of their weighting factors. A discussion on why these three risk factors have lower or no significance is given later towards the end of this section. But before that, we look into the result of the other two models in order to observe whether they are giving similar results or not.

5.7.2 Influence of risk factors from logistic regression

The binary logistic regression method models the dependent variable (i.e., the outcome) as the odd ratio against the linear sum of independent variables, each of which has a coefficient as a multiplying factor. For this analysis, the categorical variables were also converted to dummy variables before implementing the model. We ran the model in the ‘R’ programming language. The exact code to execute the model is given in the Appendix H. We report the coefficient values in the Table 5.13. The result is obtained from the execution of fold-0; other folds gave similar results.
Table 5.13: Coefficient values against risk factors in binary logistic regression model

<table>
<thead>
<tr>
<th>Risk factor</th>
<th>Coefficient value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age</td>
<td>6.20462</td>
</tr>
<tr>
<td>Sex</td>
<td>-0.14172</td>
</tr>
<tr>
<td>Behaviour</td>
<td>0.1749</td>
</tr>
<tr>
<td>Graph node match</td>
<td>27.37487</td>
</tr>
<tr>
<td>Graph pattern match</td>
<td>1.96632</td>
</tr>
<tr>
<td>Cluster match</td>
<td>0.87465</td>
</tr>
</tbody>
</table>

The table shows that the graph node match score again made the most important contribution to predicting the risk, obtaining a coefficient value of 27.37. This indicates that with each increase in the graph node match score, the log odds of developing diabetes were increased by 27.37. The second most important contribution is again attributed to the age score, with a coefficient value of 6.2. The graph pattern match score attracted the third highest weighting factor. These results are analogous with those of the parameter estimation model.

5.7.3 Influence of risk factors from binary tree classification method

The binary tree classification method sequentially splits the training set of patients into two groups based on the range condition of any of the independent variables. The resultant two groups are either assigned with a prediction outcome or are further split again based on another range condition. We implemented the binary tree classification method on the dataset using the ‘rpart’ package in ‘R’ language. The relevant code listing is given in the Appendix. For most of the iteration over 10 folds, the binary tree classification made three splits: firstly based on graph node score, secondly based on age score and finally on pattern match score. For two cases, an additional split was made on the cluster match score. Also, in three cases, there was only one split based on the graph
node match score. Figure 5.10 shows the classification tree generated from the fold-0 iteration, which has three splits.

![Classification tree generated from the binary tree classification method (output from fold-0 iteration)](image)

**Figure 5.10: Classification tree generated from the binary tree classification method (output from fold-0 iteration)**

In the figure, we can see that three splits were made based on two risk factors, i.e., graph node (gNode) and age, in that order. Graph pattern match score did not come significantly in the fold-0. This is because the other two risk factors in the first three splits gave optimal accuracy according to the tree classification algorithm. From the figure it might seem that the second and third split based on the age score can be summarised and executed as one single split, i.e., with the condition: 

\[
\text{age} < 0.3364
\]

However, the algorithm still made two successive splits on the same risk score. Actually, this is possible according to the way the binary tree classification algorithm (Breiman, et al., 1984) works. The algorithm grows the tree by progressively splitting the tree based on one risk factor at a time. In each split, it does not know in advance about the next split condition. The algorithm is based on
heuristics or optimal assumption. According to that, in each split, the algorithm chooses one risk factor and boundary condition so that the drop in impurity is maximal (along with some other conditions). Impurity (Loh, 2011) is a measure of how well the tree splits the data based on factors such as misclassification count after split etc. Several measures of impurity exist, e.g., node impurity, entropy impurity, Gini index etc. Regardless of the impurity measure, the choice of risk factor during a split is based on local criteria. Hence, the second split was based on a boundary condition on age score (i.e., $age < .5045$). And when the decision of the third split was needed, the algorithm again found age as the optimal risk score to split on. There are ways to minimise or summarise the splits after generating the tree e.g., through pruning. As our tree was not very big, we did not apply pruning and reported the result as it is.

The tree classification method outputs the relative importance of the individual parameters. For most of the cases in the 10 iterations, only three parameters were given importance. The remaining three, which are behaviour risk score, sex score and cluster match score, were not given any importance scores for most of the cases, as they were not utilised in the classification tree output. The importance scores assigned against each risk score over the 10 folds are shown in Figure 5.11.
Figure 5.11 confirms that the graph node match score consistently obtained the highest importance scores of all parameters, mostly within the range of 60–70. In all cases, this was the first parameter to split the full training set, which indicates that using this parameter at first gives the most accuracy in the primary splitting, which is later fine-tuned by the subsequent two parameters. The next parameter is age score, which received importance scores in the range of 16–26 across the folds. This was followed and often superseded by graph pattern match scores, in the range 14–23. In fact both pattern match scores and age scores often received similar importance. Overall, we can observe that the result is analogous with that of the binary logistic regression and parameter estimation models.

To summarise, in this section, we have generated a predictive model that can assess the risk of diabetes. To train the model, we used three partitions for a total of 690 diabetic patients and 690 non-diabetic patients. We used three different methods for predictive modelling, and all of them yielded relatively similar output in terms of the influence of the
six parameters or independent variables. The graph node match score received the highest influence scores, and ones significantly greater than the scores assigned to the other parameters. The age score was the second most important risk factor in predicting diabetes risk, while the graph pattern match score received relatively smaller scores.

We now proceed to discuss why the other three risk factors received lesser or no influence scores in the predictive modelling. We have observed that the male proportion was slightly higher in the diabetic cohort in the research dataset, but still it did not have much significance for the predictive modelling. We can attribute this to the fact that giving higher weighting factors or importance to the sex score would accurately predict risk for some patients, but the corresponding increase in accuracy due to this assignment would not be significant. Further, there would be a great deal of miscalculation in the forms of false positive and false negative results that would significantly reduce the accuracy. Because the predictive model looks for the best fit and it had six different parameters to balance, the sex score naturally received a lower weighting factor or importance, as there were other significant parameters that are capable of giving better accuracy.

A similar argument can be made against the very low or no significance for the behavioural risk factors, which we defined as the presence of tobacco use or smoking. Now, we know that there is scientific evidence that smoking causes type 2 diabetes (HHS, CDC, 2015, 2014): smokers are 30–40% more likely to develop type 2 diabetes than non-smokers. We may expect that this behavioural risk factor would attract some degree of importance in the predictive modelling. However, as observed, there were more significant risk scores (e.g., graph node and pattern match, age) that would outperform the accuracy results. Another crucial fact is that, although smoking seems to have a strong connection with diabetes, this does not mean that smoking does not cause any other diseases. In fact, the opposite is true. Smoking is responsible for many other diseases, including cancer and asthma. Here lies the vital clue: the administrative dataset represents the medical histories of patients, every one of which had some sorts of diseases. It is not the case that we generated the baseline network from a cohort of
diabetic patients and cohort of non-diabetic patients who were free from any smoking-related comorbidities. Rather, many of the non-diabetic patients in the baseline and training partitions within the whole cohort had cancer, asthma, and other diseases in which smoking is implicated. For this reason, despite its strong link with diabetes, smoking (a behavioural risk factor) does not attract significance in the predictive modelling for our thesis. However, as our framework is adaptive, if the quality of the dataset is different where the non-diabetic cohort does not have significant smokers or has better quality of coding (sometimes smoking is poorly coded in the hospital admission data, see limitations in 6.3.6), the predictive model should give smoking risk factor a higher weight.

We can also observe that none of the models assigned much weight or importance to the cluster match score. We actually forecasted this when we analysed the modularity structure of the baseline network in Section 5.5. The baseline network was quite saturated, as 58% of edges were present in the network. This made the modularity score lower; therefore, the relations between the comorbidities of the same cluster had little difference compared to the comorbidities of the different clusters. So, in the predictive model training phase, the cluster match scores between diabetic and non-diabetic patients did not show significant differences. Further, a large number of training patients’ cluster match scores were zero, because these patients had a small number of overall comorbidities in their admissions. Therefore, the cluster match score did not show any significance.

5.8 Risk Assessment Results for Test Group

At this stage, we apply the predictive modelling on the test cohort. There were three predictive modelling methods we implemented in the previous section. Therefore, we applied each of these separately on the test cohort. The test cohort was the only remaining partition out of the 10, after utilising the other nine partitions for the baseline network generation and training the predictive modelling. Therefore, we had 230 diabetic and 230 non-diabetic patients on whom to test the prediction results. Testing of the binary logistic regression and binary tree classification methods was done in the ‘R’
language, and testing of the parameter estimation model was implemented in our custom software (see appendix F and J for some details of the software).

We should also note that the binary logistic regression method gives its result in terms of either a log of ‘odd ratio’ or the probability (i.e., p-value). We chose to obtain the probability directly from the output against each of the test patients. To make the result comparable with other methods, we flagged the prediction result as ‘diabetic’ if the probability was greater than 50%, and ‘non-diabetic’ otherwise. In binary logistic regression, there are further provisions to adjust the threshold value by looking at the output probability with the actual risk and thus setting an optimal threshold point other than 50%. However, we skipped this adjustment, as we were already in the test phase, and the regression model was already trained. Table 5.14 shows the prediction results for the three predictive models applied on the test partition. The result was obtained from the first iteration (i.e., fold-0) of the k-fold \((k = 10)\) validation process.

<table>
<thead>
<tr>
<th></th>
<th>Binary logistic regression</th>
<th>Parameter estimation model</th>
<th>Binary tree classification</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>n=460</td>
<td>n=460</td>
<td>n=460</td>
</tr>
<tr>
<td>True positive</td>
<td>193</td>
<td>210</td>
<td>223</td>
</tr>
<tr>
<td>True negative</td>
<td>191</td>
<td>181</td>
<td>172</td>
</tr>
<tr>
<td>False positive</td>
<td>39</td>
<td>49</td>
<td>58</td>
</tr>
<tr>
<td>False negative</td>
<td>37</td>
<td>20</td>
<td>7</td>
</tr>
<tr>
<td>Accuracy (%)</td>
<td>83.48</td>
<td>85</td>
<td>85.87</td>
</tr>
</tbody>
</table>

The binary tree classification shows the highest accuracy of 85.87%, followed by the parameter estimation model, which gave 85% accuracy. Binary logistic regression performed with slightly lower accuracy than the parameter estimation model, at 83.48%. In all methods, the false positive count is greater than the false negative. This is desirable for our framework. Because our aim is to predict the risk of chronic disease (i.e., diabetes),
we can afford some false positives—that is, some patients may be flagged as having a high risk of diabetes even though they are not actually on the diabetes pathway. Eventually, the clinical diagnoses or preventive measures that would be applied for the flagged patients would reveal the false positive nature of the prediction, which would have little adverse effect on the health condition of the patient, except that some amount of clinical resources may be wasted. Conversely, having more false negatives would leave a significant number of patients undetected who are in fact on the diabetes pathway. Unmanaged and undiagnosed diabetes can lead to further complications and often to an irreversible point at which it exerts significant health, clinical and economic burdens; therefore, it is always preferable to adopt slightly aggressive margins in detecting risk in order to reduce false negatives, at the cost of having some false positives.

We can attribute the small amount of accuracy gain in our parameter estimation model over the logistic regression model to the way the result is interpreted. Although both of the models assume a linear relation, there is a small difference. The parameter estimation model gives the output as a single risk score. After obtaining all the scores from the training cohort, the model finds the optimal threshold score. All the patients scoring higher than the threshold score are flagged as diabetic, and non-diabetic otherwise. On the other hand, in the logistic regression, the linear model is generated against the log of the odd ratio. Therefore, the logistic regression reveals the increment amount of log odds ratio against any independent variables through the coefficient. Therefore, the log odds ratio increases linearly with the independent variables, not the actual probability itself. For our parameter estimation model, the risk score itself is in linear relation to the risk scores, or independent variables. This difference may affect the differences in the accuracy and contribute to the gain in favour of our parameter estimation model.

Now, we try to explain the accuracy gain of the binary tree classification model, which outperformed the other two by a slight margin. The binary tree classification is not a linear model, as discussed earlier. In fact, it is more of a data mining-based classification method, and it does not target all the independent variables at once. Rather, for our
implementation, it partitions the whole test cohort sequentially, each time against one variable with a branching condition. This gives a logical advantage and better freedom in compartmentalising the individual effect of the variables. For example, the classification method has the freedom to split the classification tree several times based on the same variables using different boundary conditions. To illustrate, one split can be made based on graph node match score, say $g_{Node} \geq .5$. Now, if $g_{Node} < .5$, it can further make another split based on the different boundary value, say $g_{Node} \geq .3$. Therefore, there would be three effective regions based on graph node match risk score, i.e., if $g_{Node} \geq .5$, if $.5 \geq g_{Node} \geq .3$ and if $g_{Node} < .3$. Thus the tree-based classification method is able to look at the different continuum of a parameter’s score, partition it into two or more regions, and predict on the class membership individually for each of the regions. This advantage is not present in either of the other two linear models.

The accuracy result from all three methods is quite satisfactory considering the complex and clinical properties of the data and the fact that the data has some practical limitations. We now discuss the inaccuracies that were observed in the models. Figure 5.12 shows the risk scores of the test cohort according to the parameter estimation model and first fold (fold-0). The red points indicate the patients who are actually diabetic, and the blue dots indicate non-diabetic patients. The threshold lines indicate the values above which all of the patient scores are predicted as diabetic and vice versa. We have discussed the four possible validation outcome of the prediction—true positive, true negative, false positive and false negative, the first two being the correct predictions. The aim of any good predictive model is to measure the overall risk score (y-axis values) and set a threshold line in such a way that all the diabetic patients (i.e., red dots) lie over the line, and all the non-diabetic patients (i.e., blue dots) lie below the line.
Figure 5.12: Overall risk scores of the test cohorts, predicted outcomes and corresponding grouping according to the parameter estimation model

From Figure 5.12, we can clearly see how the false predictions were made. Some diabetic patients (i.e., red nodes) fell under the threshold line, and were thereby flagged as false negatives, and some non-diabetic patients (i.e., blue nodes) went over the threshold line, resulting in false positives. The model cannot shift the threshold point up or down in such a way that more false positives and false negatives fall into the correct partition, because improving results for one would displace more of the other to the wrong region. The model always finds the optimum threshold yielding the highest possible accuracy given the overall risk scores. The quality of the prediction model, therefore, is mostly dependent on the definitions of the individual risk factors (e.g., age, sex, graph node match). For this research, we observed that the graph-based risk scores (e.g., node and pattern match score) and age scores were the best predictors in combination.
Now, if we look at the positions of the wrong predictions in Figure 5.12, we can clearly see that the false positive and false negative points are scattered around a narrow region near the threshold line. This indicates that these patients have very similar disease pathways (as we have a high influence of node match scores) and age scores to borderline diabetic patients, and thus their overall risk score is closer to the threshold. There is a high level of uncertainty associated with these patients, as they have a similar clinical footprint in the admissions data, but in reality may be either diabetic or non-diabetic. The model must balance their categorisations to ensure maximum accuracy. Because these results come from real-life clinical data, we argue that these false positive and false negative results scattered within a short range of the threshold line are normal and acceptable, and that the predictive model overall performed very well.

As a future step, we could modify the model by defining this short strip around which the majority of the false positive and false negative patients are scattered, instead of drawing a threshold line. Patients inside this borderline range or strip could be defined as ‘borderline patients’, whose disease risk cannot be clearly distinguished correctly, thus requiring further clinical investigation or data. With this modification, the model should give three types of prediction output: diabetic, non-diabetic and borderline. However, we leave this definition to future work, as this research has focused on creating a model that gives ‘yes’ or ‘no’ risk prediction. So far, the results yielded have been satisfactory.

5.9 Validation and Comparison of Results

In this section, we describe the process of validation of the framework in terms of accuracy, performance and consistency. We used k-fold validation with $k = 10$ as the main validation method, which we have mentioned several times throughout this chapter. Using this method, we partitioned the research cohort of 2,300 diabetic and 2,300 non-diabetic patients into 10 equal-sized partitions, referred to as fold-0, fold-1, … , fold-9. These partitions were used for three different purposes. Six of them were used for creating the baseline network, three were used for training the model and the remaining one for testing the model. We generated 10 combinations of the partitions so that in each
combination the test partition is different (see Section 4.1 for a detailed diagram) and the rest is distributed for baseline network creation and training purposes. We then implemented the full predictive modelling workflow using three methods (i.e., parameter estimation, logistic regression and binary classification) on different combinations or ‘folds’. The accuracy results of the 10 folds are given below in Figure 5.13.

![Graph showing prediction accuracy for all folds for all three methods](image)

**Figure 5.13: Prediction accuracy (%) for all folds for all three methods**

The figure shows consistent results for all the three different predictive models across the folds. The binary tree classification method shows relatively better performance than the other two in all the folds. Except for folds 4 and 5, the parameter estimation method showed slightly better performance than the binary tree classification method. As the deviation across the folds is quite small, we can safely ignore it and consider the accuracy performance consistent across the folds.

**5.9.1 Precision and recall result**

Figure 5.14 shows the precision and recall measures across all the folds for the three models. The measures are nearly consistent across the folds, as their deviation is not very high. One noticeable feature for all the measures is that the recall values are in most cases greater than the precision measures. Precision reveals the fraction of the result that is
relevant—that is, out of the all the patients who are predicted to be at high risk of diabetes, precision reveals how many are actually diabetic. Thus, if the false positive percentage is lower, it gives a higher precision result. In contrast, recall shows the fraction of relevant results that the model is returning—that is, it indicates out of all the patients who are actually T2D positive, how many the framework is correctly identifying as at high risk. Therefore, recall is inversely associated with false negatives. If the rate of false negatives is lower, it results in a higher recall value.

We observed earlier that most of the false positive and false negative values were represented by closely situated points (i.e., for the parameter estimation model they were all near the threshold line). The relation between these two types of false outcome is opposite: if the model is adjusted to bring in more false positives as true negatives, some true positives near the borderline will be shifted to false negatives, and vice versa. However, we have observed that false positives are better than false negatives from a clinical perspective. Because of their relations with precision and recall, we prefer a higher recall value (hence lower rate of false negatives) over a higher precision value. As Figure 5.14 showing this quite clearly, it indicates that the framework is performing as intended.

Figure 5.14: Precision and recall measures for all folds for all three methods
5.9.2 Effect of considering collapsed ICD codes

So far, we have considered 35 selective comorbidities, represented by grouping related ICD codes into 35 comorbidity ‘buckets’. The motivation behind this was that we only want to consider comorbidities that are clinically validated and related to diabetes, and exclude all diseases or health conditions represented by ICD codes not included in the selective list. This limits the number of comorbidities (i.e., graph nodes), and as a result, the transitions are saturated: that is, the graph density increases and network modularity decreases. To validate the effect of using these selective comorbidity ‘buckets’, we returned to the start of the framework and re-ran the whole procedure without any selective comorbidities. Instead, we used all the ICD codes in the dataset minus their suffixes (i.e., all ICD codes were shortened to three characters). We used the k-fold validation and the three predictive modelling methods in the same manner as before. Table 5.15 shows the result.

<table>
<thead>
<tr>
<th>Table 5.15: Comparison between different predictive models using collapsed ICD codes and selective comorbidities</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Binary logistic regression</strong></td>
</tr>
<tr>
<td>Accuracy (%)</td>
</tr>
<tr>
<td><strong>Collapsed ICD codes</strong></td>
</tr>
<tr>
<td><strong>Selective comorbidity</strong></td>
</tr>
</tbody>
</table>

*SD = Standard Deviation*

The table shows that using all ICD codes collapsed to three characters results in a certain amount of accuracy gain for all three predictive models (in expense of higher standard deviation, we will discuss later). This is quite an interesting result at first. In particular, the binary tree classification, which outperformed the other two methods in the previous run...
of the framework using selective comorbidities, also outperforms the others when all ICD codes are used—in fact, it does so by a greater margin, reaching 90% accuracy. Across the three methods the accuracy gain is almost consistent – about 3%-4% gain for each. In both comorbidity and selective ICD code approaches, the age, sex and behaviour risk scores remained the same, as they are individual scores and do not depend on the baseline network. Therefore, the accuracy gain must have come from the graph and cluster match scores. After examining carefully the overall individual risk scores and the way the additional accuracy was derived, we can observe a few reasons behind the accuracy gain.

The first and probably most important reason for the accuracy gain is that the selective comorbidity approach, which was derived from the Elixhauser index, has a limited number of ICD codes. Therefore, a large number of ICD codes not included in the selective comorbidity approach are eventually included in the collapsed ICD code approach. Some of these ICD codes were distinctive of the diabetes patient’s cohort, and therefore received high weightings in the baseline network via attribute adjustment, eventually resulting in higher accuracy. We can argue that the selective comorbidity condition lists only the clinically validated ICD codes and in turn comorbidities, and if all ICDs are used, there is a large number of ICDs present that may not necessarily indicate disease. However, we have observed from the accuracy gain that some of these otherwise unused ICDs in the selective comorbidity list were actually useful in predicting risk, and hence are useful to study. These ICDs may not be directly related to any disease or comorbidity, but they may indirectly point to some underlying comorbidities that are more exclusive for diabetes, and hence help in the prediction.

Another reason for the accuracy gain in the collapsed ICD approach may be that it creates a larger number of nodes in the graph. Relatedly, because individual nodes are considered, the number of transitions in the resulting overall baseline network does not saturate the whole network. This lower saturation level resulted in lower network density and higher modularity, and therefore, the number and quality of clusters are higher than in the selective comorbidity approach. This resulted in higher measures of importance for
the cluster match score in the predictive models—recall that the cluster match score was not significant in the other approaches. The improved cluster result also influenced the overall performance gain. Especially for the binary classification method, which has some advantages over the linear models in that it is better adapted to the increased number of nodes and finer clusters in the baseline network, overall performance gain was better than the other two predictive models.

The collapsed ICD code approach has some disadvantages. First, not all of the codes are clinically validated or checked. Although from a data mining and exploratory research perspective, this method is useful and allows for more accurate prediction, a clinical validation may be necessary at least to filter out some irrelevant codes to prevent overestimation. Second, we observe a higher standard deviation in the accuracy percentage for all three predictive models. Binary tree classification has a slightly lower standard deviation of 9.67. The other two models exhibited 11.95 (binary logistic regression) and 11.39 (parameter estimation model) as standard deviation. These deviations indicate that across the 10 folds of executions, collapsed ICD code approach for each of the three models showed fluctuating accuracies—significantly more than those of selective comorbidity approach. In selective comorbidity approach, the standard deviations were very nominal, i.e., within 1.23 to 1.55 range. The higher standard deviation in collapsed ICD approach indicates that due to granularity in the baseline network data, this approach may be influenced by random noise, bias and over or underestimation—subsequently resulting into fluctuation of accuracy across the 10 folds. This, in turn, pulls up the standard deviation. Therefore, although the collapsed ICD code approach shows slightly better prediction, it comes at a price of higher standard deviation. And careful consideration should be made when applying it in other contexts, that is, for other datasets or chronic diseases.

5.9.3 Attribution effect

We have claimed previously that attribute adjustment can give greater weight to the comorbidities or progressions that are more exclusive to the positive baseline network of
diabetic patients compared to the negative baseline network of non-diabetic patients. In this way, it allows for better prediction performance. In this section, we test this claim. We implemented the whole framework again using all the alternate options that we have introduced so far—using all three predictive models and both types of node format (selective comorbidity and collapsed ICD). This time, we did not use the negative baseline network to adjust the attributes of the positive baseline network. Rather, we merely normalised (scaled down) the node and edge attributes of the positive baseline network in the range of 0 to 1, and thus formed the final baseline network without attribution adjustment. All other elements of the framework implementation are executed as usual on all possible pathways. Table 5.16 below compares performance across all the different methods against the approaches with and without attribute adjustment. Also to note that we have reported the complete result set for all approaches and across all 10-folds in the Appendix I. It includes measures of true positive, true negative etc. as well as the precision, recall and accuracy.

Table 5.16: Performance comparison between inclusion and exclusion of attribute adjustment

<table>
<thead>
<tr>
<th>Attribute adjusted?</th>
<th>Node type of baseline network</th>
<th>Binary logistic regression</th>
<th>Parameter estimation model</th>
<th>Binary tree classification</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Accuracy (%)</td>
<td>SD</td>
<td>Accuracy (%)</td>
</tr>
<tr>
<td>YES</td>
<td>Collapsed ICD codes</td>
<td>85.56</td>
<td>11.95</td>
<td>88.59</td>
</tr>
<tr>
<td></td>
<td>Selective comorbidity</td>
<td>82.56</td>
<td>1.55</td>
<td>85.5</td>
</tr>
<tr>
<td>NO</td>
<td>Collapsed ICD codes</td>
<td>69.43</td>
<td>2.00</td>
<td>69.76</td>
</tr>
<tr>
<td></td>
<td>Selective comorbidity</td>
<td>78.13</td>
<td>2.24</td>
<td>85.28</td>
</tr>
</tbody>
</table>

SD = Standard Deviation
We observe from Table 5.16 that when selective comorbidities are used, the approaches with and without attribute adjustment do not differ much. At most, a performance gap of 4.43% was observed for the binary logistic regression method (82.56 vs. 78.13) without attribute adjustment. This is not very significant because in any of the two methods (selective comorbidity plus with and without attribution adjustment) we can reach 86% accuracy (in tree classification). Now, there are some reasons for selective comorbidity to have very little effect of attribute adjustment. Because we limited the number of comorbidities to some specific selections beforehand, the prevalent or more exclusive comorbidities were already included in the selection. Moreover, the negative baseline network in practice has little to change through the process of attribute adjustment. This could be understood by looking at the network properties of the negative baseline network in Table 5.8 (see Section 5.5) before. The number of edges and network density were not high compared to the positive baseline network; in fact, they were almost half of the equivalent properties for the positive baseline network. This indicates that the negative baseline network did not contribute much to the attribute adjustment. Thus, when the same method was applied without attribute adjustment, whatever adjustment effect was missed later was compensated for by the training of the predictive model. The predictive modelling always has a training phase, in which we provided an equal number of actual diabetic and non-diabetic patients to adjust the scores, thresholds and split points, and eventually succeeded in compensating for the missing attribution effect.

On the other hand, when collapsed ICD codes were used, the effect with and without using attribution adjustment made considerably bigger difference in accuracy: approximately 86% vs. 69% for logistic regression, 87% vs. 70% for parameter estimation and 90% vs. 75% for binary tree classification. We can clearly see that the method using all ICD codes did benefit a great deal from attribution adjustment, and this may be another reason why the fine-tuned graph attributes and resultant cluster information aided in the performance gain over selective comorbidity approach. As attribution plays a major role in
the collapsed ICD code approach, removing the adjustment results in imbalances in the node and edge weights to a degree, so that the individual graph- and cluster-based matching scores are infused with noise. Therefore, the predictive modelling cannot perform well. Even the best performing method in our research, the binary tree classification, reached only 78% accuracy.

Finally, we want to briefly discuss the overall advantage, disadvantage and the applicability of the framework. First of all, the methods in the framework showed promising results. Especially the selective comorbidity approach exhibited high prediction accuracy and lower standard deviation across the 10 folds. Next, we have seen that different variations of our approach revealed different functional characteristics of the prediction model and properties of the baseline network. In terms of the predictive model, binary tree classification showed the best performance across all different approaches. This makes it a likely candidate for predictive modelling based on our framework. Also, in most cases, parameter estimation method performed almost as efficiently as the binary classification method. Now, in terms of choosing between selective comorbidity and collapsed ICD code, we have seen that the former approach yielded consistent, stable and moderately accurate results (i.e., around 86% accuracy, standard deviation of around 2; see Table 5.16). On the other hand, selective comorbidity approach along with attribution adjustment yielded the highest accuracy (90%) but proved to be relatively unstable (standard deviation of 9.67 for binary tree classification, see table 5.16). Attribution adjustment did not show much performance gain when selective comorbidity was used. On the contrary, it showed considerable performance gain (around 15% gain, regardless of the type of predictive model, see Table 5.16) when collapsed ICD codes were used. We already have discussed the possible reasons in the previous paragraphs. Finally, when it comes to the choice of appropriate approach based on our framework, for any new related dataset – the choice will depend on several factors. These include, but not limited to – dataset quality, coding quality, clinical knowledge on the disease and comorbidities in question as well as research approach etc. Each of the approaches has their own advantages and disadvantages as we have discussed
earlier. Selective comorbidity might be a preferred approach if the chronic disease, comorbidities and related codes are well understood and clinically well-defined. If the research is exploratory, i.e., focused on finding major diseases and transitions that are prevalent for particular chronic disease then collapsed ICD code-based approach might be appropriate. Regarding predicting model, in most cases, binary tree classification model or parameter estimation model can be used. Because binary tree classification model is well implemented in many software packages like R, it might be easier to implement. Now, regarding attribute adjustment – it should be a preferable choice when the dataset have patients from both cohorts, i.e., one having the particular chronic disease and the other not having that chronic disease. Finally, one should include the possible socio-economic risk factors as well, whenever available in the dataset. Some of these (i.e., age etc.) had shown significance in our prediction. And, even some of them had not; they might be proved influential in the similar research using different dataset. We now summarise this chapter in the next section.

5.10 Summary

In this chapter, we have implemented our framework, presented the result and discussed the findings in broad depth. Below is a summary of the procedural steps that we have implemented in this chapter.

1. We filtered and sampled a research cohort of 2,300 diabetic and 2,300 non-diabetic patients from an administrative database of 748,982 members and 384,538 admissions.
2. We divided the research cohort into 10 equal partitions.
3. Using six partitions, we generated baseline networks for the diabetic and non-diabetic cohorts, and, through attribute adjustment, generated the overall baseline network.
4. We used three different predictive modelling techniques: two of them were linear models, and one was a data mining-based classification tree model. Three of the 10 partitions were used to train the models. We had six parameters whose significance was determined by the models.

5. The remaining partition was similarly used to predict risk with the help of the baseline network. All three predictive models were utilised, and we compared the results.

6. Steps 3 to 5 were repeated 10 times with different combinations of partitions, thereby constituting a k-fold validation.

7. Steps 3 to 6 were repeated using collapsed ICD codes as graph nodes instead of selective comorbidities, and the results were compared.

8. Steps 3 to 7 were repeated with the modification that in step 3 we did not do the attribute adjustment. The results were compared.

Following the above steps, we implemented our framework on the health dataset and presented the results from different approaches. The overall performance in terms of prediction accuracy and stability was satisfactory. We discussed the advantages, disadvantages and consideration of different approaches of the framework at the later part of this chapter. In the Table 5.17 we present a short summary of it.
**Table 5.17: Summary properties among different approaches used to implement the framework**

<table>
<thead>
<tr>
<th>Properties</th>
<th>Selective comorbidity</th>
<th>ICD code collapse</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Accuracy</strong></td>
<td>High (83% - 86%)</td>
<td>High (78% - 86%)</td>
</tr>
<tr>
<td><strong>Stability</strong></td>
<td>Very stable (SD ≈ 1)</td>
<td>Stable (SD = 2)</td>
</tr>
<tr>
<td><strong>Susceptibility to noise in dataset</strong></td>
<td>Not much</td>
<td>Not much</td>
</tr>
<tr>
<td><strong>Quality of clinical codes</strong></td>
<td>Good</td>
<td>Mostly good</td>
</tr>
<tr>
<td><strong>Computation time</strong>&lt;sup&gt;5&lt;/sup&gt;</td>
<td>Fast</td>
<td>Fastest</td>
</tr>
<tr>
<td><strong>Predictive model that performs better</strong></td>
<td>Binary tree classification, parameter estimation</td>
<td>Binary tree classification, parameter estimation</td>
</tr>
<tr>
<td><strong>Potential application area or research suitability</strong></td>
<td>When comorbidities are well-defined and clinical knowledge is available or when the user wants to investigate specific comorbidities.</td>
<td>Same as previous (left table cell), specifically when non-chronic patients' data is not available.</td>
</tr>
</tbody>
</table>

att. adj. = attribute adjustment

<sup>5</sup> Computation time is approximate and relative. May vary based on the dataset size and algorithm used.
Chapter 6: Conclusion

In this final chapter, we present the conclusion of the thesis. We start off by briefly describing what we intended to do in this thesis, how we implemented them and what we have found as the result. Next, we describe the implication of our study in different contexts. As this research deals with the real-world dataset, we faced a number of data-related issues that might affect the result to a certain degree. Therefore, in the subsequent section, we discuss the potential limitations of the study. Finally, we give future research direction for researchers to utilise, extend and improve our framework.

6.1 Overall Summary and Key Findings

With a growing burden of chronic diseases on public health and the economy, researchers and other stakeholders are increasingly being challenged to understand disease trajectories in order to design and adopt preventive policy. In this thesis, we extensively reviewed the evolution of methods in the literature that are closely aligned with this research area, and discussed the methodological and empirical gaps that led to the undertaking of this research. We focused on the issue of understanding chronic disease progression and providing an early detection framework for high-risk patients without consuming excessive clinical resources. We leveraged the existing administrative health data, more specifically hospital admission records.

We aimed to explore and answer two broad questions: (i) How can we utilise administrative data to understand chronic disease pathways? and (ii) Is it possible to reasonably predict chronic disease risk from the understanding of disease pathways? We took a network-based approach to answering these questions. A methodological framework was developed to systematically analyse the administrative data, understand the disease progression through a baseline network and then develop a predictive framework based on that baseline understanding. We also kept the framework generic so that it can be implemented on different healthcare data with minimum modification. We
further provided a rigorous discussion on the consistency and security issues associated with the administrative health data and how these issues can be addressed.

We implemented our framework on real-world datasets to test their performance and applicability and the potential of the framework from a stakeholder or business perspective. Type 2 diabetes (T2D) was chosen as the chronic disease the framework would learn to predict, and on which to test its accuracy. As per the framework, we generated a baseline network for T2D. Diabetes-associated comorbidities and their progression over time were thoroughly discussed: we demonstrated that the baseline network can reveal insights about the progression of T2D and discussed how these insights are consistent with present understandings from a clinical perspective. We utilised these empirical understandings of the comorbidities that lead to type 2 diabetes to construct a prediction model in the second part of the implementation.

The implementation further went to explore the effects of different execution pathways in terms of different methods of predictive modelling, inclusion vs. exclusion of attribute adjustment and level of ICD code groupings. This explorative analysis approach was necessary because, given the heterogeneous nature of administrative data and the existing knowledge on the progression of chronic diseases, one combination of methods might not always work well for other chronic diseases and administrative datasets. Further, we found that without attribute adjustment, the predictive framework shows relatively lower prediction accuracy. Especially, the accuracy was significantly lower when collapsed ICD code approach was used without attribute adjustment. This reveals the importance of comparing the disease trajectories of chronic patients with non-chronic patients in order to formulate a final baseline network that can properly expose the deciding factors that lead to chronic disease. Many approaches in the current literature seem to ignore the trajectory of non-chronic patients. Thus, the present research demonstrates that such neglect could create a major limitation; attribute adjustment is a major feature in our predictive modelling.
Overall, in all cases, the data mining-based binary tree classification model showed higher accuracy than the other two linear models. Also, parameter estimation model also performed quite closer to binary tree classification model in most cases. Next, we found that using collapsed ICD code approach with attribute adjustment shows the best performance, but it comes at a price of significant instability, i.e., higher standard deviation. If the same approach is used without attribute adjustment, the framework showed poor accuracy result. We discussed earlier that the collapsed ICD codes may induce noise or bias when implemented in other similar contexts. This is because the collapsing is unsupervised and, depending on the quality of coding, it can adversely affect prediction performance or stability or both. On the other hand, selective comorbidity-based grouping derived from the Elixhauser index and clinical observation is capable of maintaining the grouping process with better control and precision. In our analysis, it showed reasonable and consistent accuracy across all the predictive models when attribution adjustments were applied. Also, towards the end of the last chapter, we discussed the potential applicability of different approaches used within the framework. Generally, for the research context where clinical knowledge about the coding, diseases and comorbidities are well-defined and understood, we should use selective comorbidity. On the other hand, if they are not well-defined or understood, then collapsed ICD code approach might be more helpful. Especially, it might be suitable when the research is focused on exploring the potential diseases and patterns as risk factors. Overall, we found that despite having different accuracy results, all the approaches put high importance on network-based measures. These measures are the unique contribution of our thesis and having high importance in the prediction model thus prove the significance of network analysis in healthcare data.
6.2 Implications of the Study

6.2.1 Research implications

The incorporation of graph theory and SNA-based methods in the field of disease progression and prediction should be the primary contribution of this thesis in terms of the research implications. This network-based approach has not been widely used in this field, especially using administrative datasets. Therefore, this research is likely to contribute to the scientific community by showing the potential use of administrative data in understanding chronic disease progressions and predicting risk through the proposed framework.

Within the framework, there are three methodological aspects that constitute novel contributions from this thesis. The first is the network representation of the disease progression from the individual admission history (i.e., from administrative data) of a particular patient cohort. We have described a detailed method for constructing such a network representation for chronic and non-chronic patient cohorts in Chapter 4. The steps of converting admission records to individual disease networks with attributes and subsequently merging them into positive or negative baseline networks through statistical aggregation should be the unique contributions of this part.

The second methodological contribution is the concept of attribution adjustment between negative and positive baseline networks to form the final baseline network representing the overall chronic disease progression. The mathematical formulation of the process that was introduced to describe this part is unique to this thesis.

Finally, in the disease prediction component, we proposed three different risk factors based on graph matching. The methods to find similarity on graph node, graph pattern and social network-based cluster matching have been presented in terms of mathematical equations. We have argued that these network similarity-based risk factors can potentially compare a baseline network with a test patient’s disease network to make a prediction on whether or not the test patient will develop the chronic condition. The implementation on
T2D has shown that these risk factors, along with age score, can offer reasonable accuracy in prediction. Although some network-based risk factors such as the cluster match score did not exhibit great importance, as other risk factors did, we have suggested that this is due to the nature of the baseline network formation and experiment setup, and argued that these risk factors are likely to be useful in other setups. Overall, the outcome validates our approach using network-based modelling and prediction methods.

6.2.2 Practical implications

This research framework and corresponding experiment design was undertaken particularly with the goal to translate the outcome into knowledge for policy makers. Therefore, the research questions that we asked at the outset are aligned with the concerns of most policy makers and health funders. Administrative data, in the form of admission records, contains more or less similar information across different countries. Our framework and experiment setup is based on such typical health records, and we implemented it on a real dataset from the Australian health context, thereby making the framework practical for implementation in other contexts as well, with few modifications. For example, the disease codes that are initially input to identify the chronic patients can be changed from T2D to other chronic diseases. For example, a researcher may want to understand the progression of essential hypertension. In that case, instead of putting E.11, the ICD-10-AM code for T2D, the researcher only needs to put I10, the ICD code for essential hypertension, to make the framework generate the baseline network for that disease. Further, the data model for the research framework is designed to be stored in relational database, making it easier to change, modify or interface with the administrative data within the framework. For example, if we want to generate a baseline network in terms of codes other than the ICD codes, this can easily be done by plugging in different code tables in the database. For example, in the Australian context, CMBS codes are used to describe the services (e.g., procedures, consultation) against which the provider bills the patients. Therefore, using CMBS codes instead of ICD will create a
baseline network showing the typical services and progression through one service to another for particular chronic disease patients.

The most important practical implication of this research should be considered its potential as a software-based analytical and forecasting tool for policy makers, especially for insurance providers. Private health funds as insurance providers always strive to make better forecasts of their expenditure in coming years. Such forecasts are essential to ensure quality of service, customer satisfaction, survival in a competitive health market and maximising revenue.

![Diagram showing different groups of members having different health status for a typical health fund and their shifts over time.](image)

**Figure 6.1: Different groups of members having different health status for a typical health fund and their shifts over time**

To give a very specific example in this context, most private health funds have large numbers of members from different age groups. Given the demographics and the type of fund, in most cases, the fund has members having different health status in terms of
chronic diseases. Figure 6.1 shows a typical grouping of members based on their current health condition. The numbers are fictional, but a realistic portrayal for many health funds. The figure also shows the trajectory after say, five years, in terms of how many members are shifted from one group to another. The figure does not show the inclusion of new members with time. Now, many of these transitions are not preventable at all, although they obviously cost money to the funders. For example, many very old patients who are currently critically ill or under palliative care will eventually die in few years. Also, many of the current patients who have chronic conditions like diabetes or cancer will either deteriorate over time or their conditions will be managed through clinical interventions, screening and lifestyle management.

Managing chronic diseases, after a certain period of onset, is very costly. However, a potential window of opportunity lies in the period when the patients are borderline chronic or the disease is still in a manageable state. In numerous cases, as we have discussed earlier, progressions towards chronic disease do not involve enough symptoms to a sufficient degree that the patient becomes motivated to go for screening. When the window of opportunity to successfully intervene ends, the patient eventually ends up in the chronic disease stage, which often requires life-long management and incurs a heavy cost. As the Figure 6.1 shows, when they are at risk of chronic disease, health funds try to identify these patients whose transition are still potentially preventable. The funds encourage these borderline patients to undertake regular screening, consultation and early intervention whenever necessary. For example, in the Australian context, the Chronic Disease Management Program (CDMP) focuses on providing care coordination and self-management support to help people with chronic disease to better manage their condition and access appropriate services in order to improve health outcomes, prevent complications and reduce the need for hospitalisation (NSW Health, 2015). Many private health funds also want to encourage at-risk patients to enrol in CDMP or similar programs.

However, the challenge is to identify the high-risk chronic or soon-to-be chronic patients who would benefit from CDMP enrolment or related screening. As there are a large
number of members in a typical health fund, and the dataset is dynamically updating, it is not feasible to manually track all of them. In most cases, the analytics team looks for simple indicators, such as age, total number of hospital admissions per year and average length of stay to identify high-risk patients. These can often give erroneous or false positive rankings, as they do not consider the type of ailments the patient has had and their progression over time while calculating risk. To better predict or assess the risk or severity of chronic diseases, our framework can suitably be integrated into health fund databases. The framework can then be utilised to rank the patients easily for the risk or severity of the chronic conditions, such as diabetes, congestive heart failure, coronary heart disease, COPD or hypertension.

In fact, our implementation has already accomplished one such practical use of our framework, using a private health fund’s admission records or administrative dataset. The software was custom made following the three-tier application framework, and executed the complete framework. Therefore, it is quite easy to implement the methodology as a software or web-based tool that can leverage admission records, disease history and progression to rank patients more accurately. In consequence, this research could be translated into a valuable commercial tool for detecting high-risk chronic patients, thus enabling private funders to save costs, or greatly assist in formulating better policy around premiums, budget allocation and forecast. These can, in turn, help to provide a better quality of care to the members by early detection of chronic disease, help to manage the onset and possibly reduce hospital admissions. Further, the baseline network generated from the framework can help in identifying the set of complexities that affect quality of care, length of stay and cost. It can also suggest any patterns linking certain diseases and people of particular demographic groups or geographical areas.
6.3 Limitations of the Study

Despite our best efforts, this research has several limitations. Most of these limitations come from the fact that the dataset is a real-world healthcare dataset. Like any other real-world dataset, it has several shortcomings that cannot be fully mitigated, especially while maintaining research scope and time constraints. The major limitations of this research study are described below.

6.3.1 Absence of primary care data

Chronic disease patients are only admitted into hospitals in serious cases. Accordingly, the diagnosis trajectory will consist of disease codes that are likely to be serious enough to occur during hospitalisation. In this scenario, some diagnosis codes might be missing or have low frequency that would otherwise be included or have high frequency if primary care data were available. Also, the first point of contact for non-emergency medical needs is usually the GP. Therefore, primary care data should contain a vast amount of information for understanding chronic disease progression. However, it is practically almost impossible to obtain primary care data on a large scale like the hospital admission data. This is because primary care facilities (e.g., GPs’ chambers, community health centres) are widely distributed and numerous. It is still difficult to monitor them centrally or state-wise under common recording and reporting schemes, like the way hospitals are managed now. At present, GPs take clinical notes in different formats, using different software packages that share little or no standard format. However, with the current development of electronic health records, it should be possible to obtain primary care data in the coming years.

6.3.2 Absence of public hospital data

The healthcare data that we analysed using the framework originated from private health funds. The members present in the dataset did essentially choose private health cover. However, most of these patients were also subscribed to the public health fund, Medicare. They may also have attended public hospitals. Records for any such admissions
are not included in the dataset. This missing information may have introduced some bias in the analysis and affected its accuracy. However, there is no way to overcome this limitation except by means of data linking (see Section 6.4.1) with public hospital records, if they exist.

6.3.3 Admission time-related biases

In our framework, when generating the positive baseline network for chronic patients, we chose the admission records up to the admission when the patients are first diagnosed with the chronic condition as present in their admission record. However, we should be careful about claiming ‘first diagnoses’ of chronic disease. The time is a rough estimation of the exact occurrence. The patient may have been diagnosed sometime long before, and might have been managing the condition themselves or through primary care. Rather, the time indicates the first hospitalisation when the condition was recorded, not necessarily the actual occurrence of first diagnosis. In our implementation, the effective timeframe for the baseline network for T2D patients is likely the sum of time until the onset of T2D plus the average time after which the T2D patients were admitted in hospital and T2D is recorded first time as comorbidity.

Another possible bias may occur because of the private health cover policy. For non-emergency medical treatments, especially for procedures, private patients can choose a suitable time, as well as their doctors. Therefore, if the elected time varies considerably, it does not show accurate time of the comorbidity onset. This can incur bias.

6.3.4 Possible bias in the number of admissions

Some admissions for the members may be missing from the dataset due to their membership plan. The healthcare funds offer different types of cover for their members. While most members subscribe to the typical health cover, some may choose extra options like allied health cover (e.g., for dentistry, ophthalmology). Therefore, some hospital admissions may occur if the patient has a more extensive health cover. On the other hand, some admissions may not occur if the patient has a more basic health cover.
Alternatively, they can be admitted without cover, and in that case the admission will not be reported to the respective health fund. Further, the patients can switch between different packages within the same health fund according to their requirements. These may introduce bias in the number of admissions.

6.3.5 Suspension of membership

There is a small chance of bias due to the fact that the members have liberty to suspend their membership. This can create bias in two ways. First, the member can terminate permanently and can choose a different private health fund, or no fund at all. Second, the member can suspend temporarily, and may choose other fund. After some period of time, they can resume their membership. This first event will not incur any bias in our analysis because, if they terminate the membership, our framework will only consider the progression up to the time of last admission. In contrast, for the second case, during the temporary suspension period, the framework will find no admission. However, upon resuming, if they are admitted again, that would be registered in the progression history. The framework will consider the disease progression from the last admission before suspension to the first admission after resuming the membership. This would incur a bias. Although, upon discussion with the data custodian, we found that the amount of such suspensions should be very small, it is still worth noting here. This type of bias also depends on the dataset design. For our case, it was not possible to filter out patients who had had suspensions or not, mostly due to the strict de-identification routines implemented during data transfer. However, other administrative datasets may have sufficient membership information to filter out such inconsistent entries beforehand.

6.3.6 Quality of coding

We discussed earlier in Chapter 3 the issue of coding quality during the hospital admissions. Here we briefly recall the potentially associated bias due to this. First, healthcare policy changes frequently in terms of coding practice. Coding related to chronic diseases as primary or secondary comorbidity has undergone several policy updates
during the last decade. These caused the frequency of related ICD codes to surge or drop dramatically within months. Although most of our data was collected outside the major policy change period, a small part remains affected by this period of change. Second, the expertise of the individual clinical coder also affects the coding quality. Normally, trained healthcare professionals record the administrative data. The process can be affected by several factors, including experience, expertise, training, workload or reference-related resources. Several other factors, like the version of the coding software (e.g., DRG grouper), motivation and time constraints can also play a role in dictating the level of coding. Finally, coding standards (e.g., version, implementation) and funding for clinical coding can also play a role.

Poor coding can result in fewer recorded comorbidities or behavioural risk factors (e.g., smoking), incorrect entries or sometimes over-coding. This can affect the accuracy of the model. For example, we manually did a random check on the false negative cases obtained from our models. It was found that, in most cases, they had a fewer number of comorbidities recorded during their admission histories. As a result, they obtained lower network- and cluster-based scores, which in turn, put them below the threshold value and identify as non-diabetic, i.e., false negative. While this could be purely by chance, the lack of comorbidities may also due to poor coding and may be worthwhile to investigate further. Now, in terms of policy and coding versions, as our research data is quite new and uses the latest ICD version, therefore, bias related to these should be minimal. Also, private funders usually cross-check the claims that come through service providers and patients during admission, along with the ICD codes in the admission summary. Claims are made against particular medical services, which are coded as CMBS codes. The matching process between CMBS codes and ICD codes can give an approximate indication of any coding inconsistency, over-claiming or fraud. In fact, fraud detection is another big research topic related to health funds. Within our research scope, the comorbidity coding, as checked by the funders, did not show a great deal of bias in terms of quality.
6.4 Future Research Directions

During the research, we faced several challenges, mostly in terms of data quality and properties. A few methodological questions and opportunities also arose, which we aimed to explore. However, as there is always an inherent trade-off between available time and features to include, we needed to set a scope for the research. It is hoped that investigation of related features beyond our present scope will be undertaken in future by us or the research community. The major research directions based on the current framework are briefly discussed here.

6.4.1 Data linkage

The research data that we used has a sufficiently large population, as it was acquired from one of the larger conglomerates of private healthcare funds in Australia. However, this is not the only health fund or data source; there are numerous private health funds, as well as the government healthcare system, Medicare, which covers healthcare services for the remaining population. One of the major reasons we needed to do rigorous filtering to choose patients is because not all patients had sufficient admission information within our research dataset. One of the reasons for this is that those patients may have terminated their membership and opted for other funds. Also, in some cases, members can temporarily subscribe to another fund for an interim period, and then again return to the original fund. Although absence for an interim period is very unlikely for our dataset, we still filtered out a number of patients who did not have sufficient information, some of whom may have opted for other funds. It may be worthwhile to track these patients. From a policy-level perspective for insurance providers, it would be interesting to see which particular patients have opted out, at what health status they did so and what policies could be adopted to keep these valuable customers with their funds.

Further, our dataset only represents around five years of data, as the framework is focused on understanding disease progression over time. Therefore, if we want to track previous or subsequent medical history for these patients, we would need to obtain their
other datasets and merge or link them. This process is called ‘data linkage’, ‘record linkage’ or ‘data integration’. Linkage between healthcare datasets involves several considerations, including privacy issues, ethical and legislative guidelines and regulations. Further, both datasets should be matched based on unique identifiers such as Medicare number. If no definitive identifier is found, a combination of identifiers, such as name, address and age, can be used to identify patients over different datasets and merge their records. This process of data linkage is often executed by third parties or government-approved organisations for legislative reasons, as data linkage needs to expose personal identifiers during calculation.

6.4.2 Inclusion of primary care data

As discussed in the limitations section (see Section 6.3), the early treatment or diagnosis of chronic disease may be likely to occur through primary care by a GP. People normally go to GP as the first point of contact in non-emergency situations. Clinical notes made by GPs and their corresponding diagnostic results are extremely valuable resources. Unfortunately, on a large scale, these data are not collected centrally using a standard format or reporting scheme. There are still considerable amounts of primary care data from GPs and self-reported data kept by stakeholders such as state health departments for the Australian context. Incorporating those forms of data into the framework, either by linking them to existing data, if possible, or simply using the primary care data as a standalone set, could be an interesting research direction. Further, in the coming years, the Australian government will roll out electronic health records for patients, which will contain patients’ entire medical history, including medications and consultation information, stored centrally and accessible from everywhere. This huge and complete health database should open up new frontiers of research, especially in understanding disease progression and prediction-related topics, in which our framework can be extended easily.
6.4.3 Effect of the lookback period

The lookback period, related to understanding disease pathways, is the amount of time before the onset of disease within which we need to look at the medical history of the patients to understand how these comorbidities eventually led to the onset. This period is often set between one and five years, depending on factors such as population size and disease types. In our case, the data itself represented a five-year time period. Therefore, our lookback period had a maximum of five years. However, the lookback period was not uniform for all patients, as not all of them had first and last admissions dates five years apart. To maintain uniformity, we normalised the admission records by time duration. Therefore, admissions with smaller time gaps between first and last admissions were scaled up to five years’ lookback period in our experiment. It would be interesting to see whether choosing admissions for one year only before the first diagnosis date could offer different degrees of accuracy or understanding of the progression. This could be done using the existing dataset by filtering the patients who have at least one year of information before first diagnosis and then using only the admission for that year.

6.4.4 Extending the framework to the post-diagnosis phase

In the first part of our framework, we generated two baseline networks, one for chronic and one for non-chronic patients, and then merged them to create the final baseline network through attribute adjustment. While constructing the positive baseline network from the chronic patients, we only considered admission records up to the point when the first chronic disease-related ICD code appeared in the admission. Therefore, we examined the disease progression prior to developing chronic disease through the final baseline network, which exposed the comorbidities and patterns that led to the onset. Keeping the same principles for the baseline network, we can alter the timeline of progression to include the post-diagnosis phase. We can consider admissions after the first chronic code appeared. When the positive baseline network is generated using this timeline, we can see what diseases and transitions are more prevalent after the chronic condition develops in patients. The negative baseline network of non-chronic patients should be kept the
same. We can then produce the final baseline network through attribute adjustment, revealing the potential diseases and transition patterns that occur more exclusively in the chronic patients after they are diagnosed with the disease. This experiment can easily be done using the existing dataset that we employed in this project.
References


40. Burgess, R.L. Communication Networks and Behavioral Consequences! Human Relations, 22 (2). 137-159.

41. Burt, R.S. The social structure of competition. Networks and organizations: Structure, form, and action. 57-91.


63. Espadaler, J., Aragüés, R., Eswar, N., Marti-Renom, M.A., Querol, E., Avilés, F.X., Sali, A. and Oliva, B. Detecting remotely related proteins by their interactions and

64. Fears, J.R. The plague under Marcus Aurelius and the decline and fall of the Roman Empire. *Infectious disease clinics of North America, 18* (1). 65-77.


68. Garland, A., Fransoo, R., Olafson, K., Ramsey, C., Yogendren, M., Chateau, D. and McGowan, K. *The epidemiology and outcomes of critical illness in Manitoba.* University of Manitoba, Faculty of Medicine, Department of Community Health Sciences, 2012.


76. Heider, F. The psychology of interpersonal relations. Psychology Press, 2013.


144. Sampson, S.F. A novitiate in a period of change: An experimental and case study of social relationships, Cornell University, September, 1968.


156. Smith Sidney, L. Communication pattern and the adaptability of task-oriented groups: an experimental study. *Cambridge, MA: Group Networks Laboratory, Research Laboratory of Electronics, Massachusetts Institute of Technology.*


166. Uddin, S. Social networks enabled coordination performance model for patient hospitalization. Faculty of Engineering and IT, The University of Sydney, Sydney, 2011.


168. Uddin, S., Khan, A. and Piraveenan, M. Administrative claim data to learn about effective healthcare collaboration and coordination through social network Hawaii
International Conference on System Sciences (HICSS-48), Hawaii, United States, 2015.


Appendices

Appendix A: Glossary

**Admission:** The formal process of undergoing treatment in a hospital. See also – *episode*, *separation*.

**Australian Classification of Health Interventions (ACHI):** A seven digit coding system for classifying interventions that has been used since 1998. It is based on MBS and was previously known as the Medicare Benefits Schedule-Extended (MBS-E). The first five digits of ACHI code indicate the MBS item number (if exists). The two-digit extension represents specific interventions included in that item.

**Australian Refined Diagnosis Related Group (AR-DRG):** Australian admitted patient classification system which provides a clinically meaningful way of relating the number and type of patients treated in a hospital (i.e., *casemix*) to the resources required by the hospital. Each AR-DRG represents a class of patients with similar clinical conditions requiring similar hospital services. (source: AIHW).

**Baseline Network:** A network representation of the *health trajectory* of chronic disease or any particular group of patients.

**Bound:** The maximum (upper bound) or minimum (lower bound) possible value of a mathematical function (or equation) for any valid input.

**Care Type:** The overall nature of a clinical service provided to an admitted patient during an *episode* of care. Examples are – acute care, palliative care, psychogeriatric care etc.

**Casemix:** The concept of providing healthcare industry with a consistent method of classifying types of patients, their treatment and associated costs. See also – *AR-DRG*.
**Chronic disease:** A broad set of illnesses and health conditions that are non-communicable and persist over a long period of time, such as – diabetes, cancer, cardiovascular diseases etc.

**Comorbidity:** Diseases that occur together. In clinical coding context, the disease chiefly responsible for the *admission* is called *principal diagnosis*. Other present diseases or conditions are recorded as secondary diagnoses and are called comorbidities. See also – *principal diagnosis*.

**Cream skimming:** (For insurance providers) Attract people with low additional or healthcare risk.

**Data linkage:** The bringing together (linking) of information from two or more different data sources that are believed to relate to the same entity—for example, the same individual or the same institution. This can provide more information about the entity and, in certain cases, can provide a time sequence, helping to tell a story, show 'pathways' and perhaps unravel cause and effect. The term is used synonymously with 'data integration' and 'record linkage.'

**Disease progression:** The occurrence of different diseases of comorbidities over time and the transitions between them. See also – *health trajectory*.

**Episode:** An episode (METeOR ID 491557) is the period between *admission* and *separation* that a person spends in one hospital, and includes leave periods not exceeding seven days. Admission and separation can be either formal or statistical. For our context, it is statistical, meaning that each episode refers to only one *care type* for clarity in reporting and resource allocation. If a patient’s care type changes during a hospital stay, such as, if transferred from acute care to palliative care, it is statistically separated into two episodes.

**Health Facility:** A place where healthcare is provided, such as - hospitals, clinics, doctor’s chambers etc.
Health Fund: In Australian healthcare context, private health insurance is provided through organisations registered under the *Private Health Insurance Act 2007*. The registered organisations are called *Health Fund*. A list of health funds operating in Australia can be found at [www.privatehealth.gov.au/dynamic/healthfundlist.aspx](http://www.privatehealth.gov.au/dynamic/healthfundlist.aspx)

Health trajectory: Evolution of the health status of a single or group of patients over time. In this research context, we used diseases (i.e., ICD codes) as the indicator of health status or health trajectory. See also - *disease progression*.

Individual disease network: The network representation of the *disease progression* of an individual using hospital admission records.

International Classification of Diseases (ICD): Disease classification system published by the World Health Organisation (WHO) for worldwide use in translating the narrative descriptions of diseases, injuries and procedures contained in medical records into alphanumeric codes.

Medicare (Australia): The universal health care scheme in Australia providing access to medical and hospital services for all Australian residents and certain categories of visitors to Australia.

Medicare Benefits Schedule (MBS): Listing of the *Medicare* services subsidised by the Australian Government. The listing is available at MBS Online website (http://www.mbsonline.gov.au/). See also – *Medicare*.

Pharmaceutical Benefits Scheme (PBS): The Australian Government program to provide subsidised medicines to Australian residents and certain categories of visitors to Australia. The schedule can be accessed from PBS website (http://www.pbs.gov.au/). See also – *Medicare Benefits Schedule*.

Primary key: An identifier that is unique for each record in a database system such as driver’s license, vehicle registration number etc.
**Principal diagnosis:** (METeOR ID: 361034) The diagnosis established after study to be chiefly responsible for occasioning an *episode* of care.

**Providers:** An entity who provides the healthcare service. These include individual GPs, specialists (doctors, surgeons, anaesthetists etc.), nurses etc. In generic sense, healthcare facilities, i.e., hospitals, clinics etc. can also be a provider.

**Relapse:** A recurrence of illness or symptoms of disease after a period of remission or improvement. See also – *remission*.

**Remission:** Absence of signs or symptoms for a chronic disease or illness after the onset, as well as the time period during which such absence remains. A disease in remission state does not indicate completely cure. See also – *relapse*.

**Separation:** (METeOR ID 327268) The process by which an episode of care for an admitted patient ceases. It can be formal or statistical. For our research context it is statistical meaning that it indicates the cessation of an episode of care for a patient within the one hospital stay.

**Simple Graph:** An unweighted and undirected graph, i.e., edges do not have any directionality or weight. Also there is no self-loop, i.e., none of the nodes have edge that points to itself.

**Universal Health Care:** Health system that provides healthcare and financial protection to the citizens of a particular country. Healthcare services under universal health care are often subsidised by the government. See also – *Medicare.*
### Appendix B: Acronyms

<table>
<thead>
<tr>
<th>Acronym</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>ACHI</td>
<td>The Australian Classification of Health Interventions</td>
</tr>
<tr>
<td>ACS</td>
<td>Australian Coding Standards</td>
</tr>
<tr>
<td>AIHW</td>
<td>Australian Institute of Health and Welfare</td>
</tr>
<tr>
<td>AMA</td>
<td>Australian Medical Association</td>
</tr>
<tr>
<td>ANOVA</td>
<td>Analysis of variance</td>
</tr>
<tr>
<td>AR-DRG</td>
<td>Australian Refined Diagnosis Related Group</td>
</tr>
<tr>
<td>BMI</td>
<td>Body Mass Index</td>
</tr>
<tr>
<td>BN</td>
<td>Baseline Network</td>
</tr>
<tr>
<td>CDMP</td>
<td>Chronic Disease Management Program</td>
</tr>
<tr>
<td>CMBS</td>
<td>Commonwealth Medical Benefits Scheme</td>
</tr>
<tr>
<td>COPD</td>
<td>Chronic Obstructive Pulmonary Disease</td>
</tr>
<tr>
<td>DNA</td>
<td>DeoxyriboNucleic Acid</td>
</tr>
<tr>
<td>DRG</td>
<td>Diagnosis Related Group</td>
</tr>
<tr>
<td>E-R model</td>
<td>Entity-Relationship model</td>
</tr>
<tr>
<td>GP</td>
<td>General Practitioner</td>
</tr>
<tr>
<td>HAMBS</td>
<td>Hospital and Medical Benefits System, a not-for-profit company formed by Australian private health insurers to meet the present and future technological demands for its member funds.</td>
</tr>
<tr>
<td>HCP</td>
<td>Hospital Casemix Protocol</td>
</tr>
<tr>
<td>Acronym</td>
<td>Full Form</td>
</tr>
<tr>
<td>---------</td>
<td>-----------</td>
</tr>
<tr>
<td>HRCN</td>
<td>Hospital-rehab coordination network</td>
</tr>
<tr>
<td>ICD</td>
<td>International Classification of Diseases</td>
</tr>
<tr>
<td>ICU</td>
<td>Intensive care unit</td>
</tr>
<tr>
<td>IDE</td>
<td>integrated development environment</td>
</tr>
<tr>
<td>MBS</td>
<td>Medicare Benefits Schedule</td>
</tr>
<tr>
<td>MPM</td>
<td>Mortality Protection Model</td>
</tr>
<tr>
<td>NCCC</td>
<td>The National Casemix and Classification Centre</td>
</tr>
<tr>
<td>NDSS</td>
<td>National Diabetes Services Scheme</td>
</tr>
<tr>
<td>NSW</td>
<td>New South Wales</td>
</tr>
<tr>
<td>PBS</td>
<td>Pharmaceutical Benefits Scheme</td>
</tr>
<tr>
<td>PCCN</td>
<td>Patient-centric care coordination network</td>
</tr>
<tr>
<td>PCN</td>
<td>Physician collaboration network</td>
</tr>
<tr>
<td>RTM</td>
<td>Released To Manufacturing</td>
</tr>
<tr>
<td>SAPS</td>
<td>Simplified Acute Physiology Score</td>
</tr>
<tr>
<td>SN</td>
<td>Social Network.</td>
</tr>
<tr>
<td>SNA</td>
<td>Social Network Analysis</td>
</tr>
<tr>
<td>T2D</td>
<td>Type 2 Diabetes</td>
</tr>
<tr>
<td>US</td>
<td>United States</td>
</tr>
<tr>
<td>WHO</td>
<td>World Health Organization</td>
</tr>
</tbody>
</table>
## Appendix C: Mathematical notation used

<table>
<thead>
<tr>
<th>Notation</th>
<th>Description</th>
<th>Explanation, bounds and possible values</th>
</tr>
</thead>
<tbody>
<tr>
<td>$N_P$</td>
<td>Individual patient’s disease network</td>
<td></td>
</tr>
<tr>
<td>$N_{test}$</td>
<td>Test patient’s disease network</td>
<td></td>
</tr>
<tr>
<td>$N_{B+ve}$</td>
<td>Baseline Network of T2D patient cohorts</td>
<td></td>
</tr>
<tr>
<td>$N_{B-ve}$</td>
<td>Baseline Network of non-T2D patient cohorts</td>
<td></td>
</tr>
<tr>
<td>$N_B$</td>
<td>Baseline Network derived from $N_{B+ve}$ and $N_{B-ve}$</td>
<td></td>
</tr>
<tr>
<td>$G(V,E)$</td>
<td>Any generic graph G where V is the set of nodes (i.e., vertices) and E is the set of Edges.</td>
<td></td>
</tr>
<tr>
<td>$V(G)$</td>
<td>Set of all vertices (nodes) of a generic graph G</td>
<td>Alternately, can be just written as $V$</td>
</tr>
<tr>
<td>$E(G)$</td>
<td>Set of all edges of a generic graph G</td>
<td>Alternately, can be just written as $E$</td>
</tr>
<tr>
<td>$</td>
<td>S</td>
<td>$</td>
</tr>
<tr>
<td>$n_{total}$</td>
<td>Number of all patients.</td>
<td></td>
</tr>
<tr>
<td>$n$</td>
<td>Number of all filtered patients.</td>
<td>$n \leq n_{total}$</td>
</tr>
<tr>
<td>$n_{+ve}$</td>
<td>Number of all filtered patients who are type-2 diabetic.</td>
<td></td>
</tr>
<tr>
<td>$n_{-ve}$</td>
<td>Number of all filtered patients who are not type-2 diabetic.</td>
<td>$n = n_{+ve} + n_{-ve}$</td>
</tr>
<tr>
<td>$C_{T2D}$</td>
<td>Cohort of all type 2 diabetic (T2D) patients</td>
<td></td>
</tr>
<tr>
<td>$C'_{T2D}$</td>
<td>Cohort of all non-T2D patients</td>
<td></td>
</tr>
<tr>
<td>$C_i(T2D)$</td>
<td>$i^{th}$ cohort of T2D patients.</td>
<td>$i = {1,2,\ldots,10}$</td>
</tr>
<tr>
<td>$C'_i(T2D)$</td>
<td>$i^{th}$ cohort of non-T2D patients.</td>
<td>$i = {1,2,\ldots,10}$</td>
</tr>
<tr>
<td>$d_i$</td>
<td>$i^{th}$ comorbidity for the selective comorbidity approach</td>
<td>$i = {1,2,\ldots,35}$</td>
</tr>
<tr>
<td>$m$</td>
<td>Total number of comorbidities that are considered.</td>
<td>In our thesis, $m = 35$</td>
</tr>
<tr>
<td>$a_i$</td>
<td>$i^{th}$ Admission for a patient</td>
<td></td>
</tr>
<tr>
<td>( D )</td>
<td>Set of comorbidities</td>
<td>( D = {d_1, d_2, ..., d_m} )</td>
</tr>
<tr>
<td>( D_{a_i} )</td>
<td>Set of diseases or comorbidities that are recorded during admission ( a_i )</td>
<td></td>
</tr>
<tr>
<td>( rf )</td>
<td>Variable to indicate risk factor</td>
<td>( rf = {rf_{age}, rf_{sex}, rf_{behave}, rf_{node}} )</td>
</tr>
<tr>
<td>( sf_{rf_i} )</td>
<td>Scaling or weighting factor for a risk factor ( i ).</td>
<td></td>
</tr>
<tr>
<td>( S_i )</td>
<td>Overall risk score of a patient ( i )</td>
<td></td>
</tr>
</tbody>
</table>
Appendix D: Letter of Human Ethics approval

Research Integrity
Human Research Ethics Committee

Friday, 20 November 2015

Dr Mohammed Shahadat Uddin
Civil Engineering, Faculty of Engineering and Information Technologies
Email: shahadat.uddin@sydney.edu.au

Dear Mohammed Shahadat

I am pleased to inform you that the University of Sydney Human Research Ethics Committee (HREC) has approved your project entitled "Identifying high-risk patterns in chronic disease progression network from healthcare data - a framework based on graph theory and social network analysis."

Details of the approval are as follows:

Project No.: 2015/824
Approval Date: 12 November 2015
First Annual Report Due: 12 November 2016
Authorised Personnel: Uddin Mohammed Shahadat; Khan Mohammad Ariful Alam; Srinivasan Uma;

HREC approval is valid for four (4) years from the approval date stated in this letter and is granted pending the following conditions being met:

Conditions of Approval

• Continuing compliance with the National Statement on Ethical Conduct in Research Involving Humans.
• Provision of an annual report on this research to the Human Research Ethics Committee from the approval date and at the completion of the study. Failure to submit reports will result in withdrawal of ethics approval for the project.
• All serious and unexpected adverse events should be reported to the HREC within 72 hours.
• All unforeseen events that might affect continued ethical acceptability of the project should be reported to the HREC as soon as possible.
• Any changes to the project including changes to research personnel must be approved by the HREC before the research project can proceed.
• Note that for student research projects, a copy of this letter must be included in the candidate’s thesis.

Chief Investigator / Supervisor’s responsibilities:

Research Integrity
Research Portfolio
Level 6, Jane Foss Russell
The University of Sydney
NSW 2006 Australia

T +61 2 8627 8111
F +61 2 8627 8177
E r.researchethics@sydney.edu.au
sydney.edu.au
Letter of Human Ethics approval (continued)

1. You must retain copies of all signed Consent Forms (if applicable) and provide these to the HREC on request.

2. It is your responsibility to provide a copy of this letter to any internal/external granting agencies if requested.

Please do not hesitate to contact Research Integrity (Human Ethics) should you require further information or clarification.

Yours sincerely

[Signature]

Professor Glen Davis
Chair
Human Research Ethics Committee

This HREC is constituted and operates in accordance with the National Health and Medical Research Council’s (NHMRC) National Statement on Ethical Conduct in Human Research (2007), NHMRC and Universities Australia Australian Code for the Responsible Conduct of Research (2007) and the CPMP/ICH Note for Guidance on Good Clinical Practice.
Appendix E: Actual E-R diagram used in the model

In chapter 3, we gave an overview of the functional properties and specification of the dataset and entities within it that are related to the framework. We did not mention the actual entity and field names that are actually used in the analysis. There are also some entities that are used for as supplementary in the analysis. Following two figures show database design (E-R diagram) that was used by the software for analysis.
### Appendix F: Technical details of the software and analytical tools used

<table>
<thead>
<tr>
<th>Purpose</th>
<th>Software and version</th>
<th>Comments</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Analysis</strong></td>
<td>Chronic Disease Risk Prediction Framework (CDRAF) v 3.1</td>
<td>The core of the framework. Used to prepare dataset, implement the methods, analyses and the prediction. We developed the software from scratch.</td>
</tr>
<tr>
<td></td>
<td>Microsoft Visual Studio (Enterprise 2015)  v 14.0.24720.00 Update 1</td>
<td>The CDRAF software was developed using this Integrated Development Environment (IDE). Notable tools were – Visual C# 2015 (main programming language) and .Net Framework (version</td>
</tr>
<tr>
<td></td>
<td>Visual C# 2015</td>
<td>Integrated within Visual Studio IDE. It was the programming language used to develop the core (CDRAF) software.</td>
</tr>
<tr>
<td></td>
<td>R Studio</td>
<td>IDE for R programming. Used for implementing two predictive models and few graphs.</td>
</tr>
<tr>
<td></td>
<td>Microsoft Excel</td>
<td>A significant part of intermediate calculation and data interfacing were done in MS Excel</td>
</tr>
<tr>
<td><strong>Software library/pack age</strong></td>
<td>ZedGraph v 5.1.7</td>
<td>Planned for dynamic visualisation. Later decided to keep out of our research scope.</td>
</tr>
<tr>
<td></td>
<td>Npgsql v 3.1.0-alpha6</td>
<td>Used for interfacing between PostgreSQL database and c# programming language with .NET environment</td>
</tr>
<tr>
<td></td>
<td>NuGet Package Manager v 3.3.0</td>
<td>Integrated with Visual Studio IDE.</td>
</tr>
<tr>
<td><strong>Database engine</strong></td>
<td>Microsoft SQL server v 12.00.2000 (SQL Server 2014 RTM)</td>
<td>The research database was hosted in this engine. The main CDRAF software communicated with the engine to perform data read and write into the database.</td>
</tr>
<tr>
<td></td>
<td>PostgreSQL</td>
<td>The originally received database was hosted in this engine. The main CDRAF software transferred the required data to our research database in MS SQL Server during the data preparation process.</td>
</tr>
<tr>
<td>Purpose</td>
<td>Software and version</td>
<td>Comments</td>
</tr>
<tr>
<td>------------</td>
<td>-------------------------------</td>
<td>--------------------------------------------------------------------------</td>
</tr>
<tr>
<td>Database client</td>
<td>SQL Server Data Tools v 14.0.50616.0</td>
<td>This database management tool is integrated within Visual Studio IDE.</td>
</tr>
<tr>
<td></td>
<td>pgAdmin v 1.20.0</td>
<td>Used for administering PostgreSQL</td>
</tr>
<tr>
<td>Visualisation</td>
<td>Gephi v 0.8.2</td>
<td>Baseline network was visualised using Gephi, a powerful Social Network Analysis (SNA) software. We also used it for disease cluster detection.</td>
</tr>
<tr>
<td></td>
<td>Adobe Photoshop CS6 v 13.0 (Extended)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Microsoft office suit</td>
<td>Mainly MS Power Point is used.</td>
</tr>
</tbody>
</table>
### Appendix G: ICD-10-AM codes for identifying selective comorbidities

<table>
<thead>
<tr>
<th>S/L</th>
<th>Comorbidities</th>
<th>ICD-10 codes</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Cardiac arrhythmias</td>
<td>I441, I442, I443, I456, I459, I477, I487, I497, RO00, RO01, RO0 8, T821, Z450, Z950</td>
</tr>
<tr>
<td>2</td>
<td>Valvular disease</td>
<td>A520, I05?, I06?, I07?, I08?, I091, I098, I34?, I35?, I36?, I37?, I38</td>
</tr>
<tr>
<td>3</td>
<td>Pulmonary circulation Disorders</td>
<td>I26?, I27?, I280, I288, I289</td>
</tr>
<tr>
<td>4</td>
<td>Peripheral vascular disorders</td>
<td>I70?, I71?, I731, I738, I739, I771, I790, I792, K551, K558, K559, Z 958, Z959</td>
</tr>
<tr>
<td>5</td>
<td>Hypertension, uncomplicated</td>
<td>I10?</td>
</tr>
<tr>
<td>6</td>
<td>Hypertension, complicated</td>
<td>I11?, I12?, I13?, I15?</td>
</tr>
<tr>
<td>7</td>
<td>Paralysis</td>
<td>G041, G114, G801, G802, G81?, G82?, G830, G831, G832, G83 3, G834, G839</td>
</tr>
<tr>
<td>10</td>
<td>Hypothyroidism</td>
<td>E00?, E01?, E02?, E03?, E890</td>
</tr>
<tr>
<td>13</td>
<td>Peptic ulcer disease excluding bleeding</td>
<td>K257, K259, K267, K269, K277, K279, K287, K289</td>
</tr>
<tr>
<td>14</td>
<td>AIDS/HIV</td>
<td>B20?, B21?, B22?, B24?</td>
</tr>
</tbody>
</table>

*For blue coloured comorbidities, some new ICD codes were added after inspection. The last 6 orange coloured comorbidities were added as addition to the original Elixhauser comorbidity index.

*?* indicates wildcard mask that matches with any character. The dot (.) in ICD code is omitted.

For details, refer to the Section 5.4.1*
<table>
<thead>
<tr>
<th></th>
<th>Disease</th>
<th>ICD Codes</th>
</tr>
</thead>
<tbody>
<tr>
<td>15</td>
<td>Lymphoma</td>
<td>C81?,C82?,C83?,C84?,C85?,C88?,C96?,C900,C902</td>
</tr>
<tr>
<td>16</td>
<td>Metastatic cancer</td>
<td>C77?,C78?,C79?,C80?</td>
</tr>
<tr>
<td>18</td>
<td>Rheumatoid arthritis/collagen vascular diseases</td>
<td>L940,L941,L943,M05?,M06?,M08?,M120,M123,M30?,M310,M311,M312,M313,M32?,M33?,M34?,M35?,M45?,M461,M468,M469</td>
</tr>
<tr>
<td>19</td>
<td>Coagulopathy</td>
<td>D65,D66,D67,D68?,D691,D693,D694,D695,D696</td>
</tr>
<tr>
<td>20</td>
<td>Obesity</td>
<td>E66?</td>
</tr>
<tr>
<td>21</td>
<td>Weight loss</td>
<td>E40?,E41?,E42?,E43?,E44?,E45?,E46?,R634,R64</td>
</tr>
<tr>
<td>22</td>
<td>Fluid and electrolyte disorders</td>
<td>E222,E86?,E87?</td>
</tr>
<tr>
<td>23</td>
<td>Blood loss anaemia</td>
<td>D500</td>
</tr>
<tr>
<td>24</td>
<td>Deficiency anaemia</td>
<td>D508,D509,D51?,D52?,D53?</td>
</tr>
<tr>
<td>25</td>
<td>Alcohol abuse</td>
<td>F10,E52,G621,l426,K292,K700,K703,K709,T51?,Z502,Z714,Z721,F102</td>
</tr>
<tr>
<td>28</td>
<td>Depression</td>
<td>F204,F313,F314,F315,F32?,F33?,F341,F412,F432</td>
</tr>
<tr>
<td>29</td>
<td>Cataract</td>
<td>H25?,H26?</td>
</tr>
<tr>
<td>30</td>
<td>Anaemia, unspecified</td>
<td>D649</td>
</tr>
<tr>
<td>31</td>
<td>Prediabetes medication</td>
<td>Z9222</td>
</tr>
<tr>
<td>32</td>
<td>Macular degeneration</td>
<td>H353</td>
</tr>
<tr>
<td>33</td>
<td>Presence of coronary angioplasty implant and graft</td>
<td>Z955</td>
</tr>
<tr>
<td>34</td>
<td>Presence of aortocoronary bypass graft</td>
<td>Z951</td>
</tr>
</tbody>
</table>
Appendix H: R commands and functions used

**R code for Binary Logistic Regression training:**

```r
fold0LogitTrainData <- read.csv("~/directoryLocation/fold0Train.txt")
fold0LogitTrain <- glm(formula=actualRisk~.,family="binomial",data= fold0LogitTrainData)
summary(fold0LogitTrain)
```

**R code for Binary Logistic Regression test:**

```r
fold0LogitTestData <- read.csv("~/directoryLocation/fold0Test.txt")
fold0LogitTest <- predict.glm(fold0LogitTrain,newdata = fold0LogitTestData[,-7],type="response")
fix(fold0LogitTest)
fix(fold0LogitTest)
```

**R code for Binary Tree Classification training:**

```r
library(rpart)
fold0TreeTrainData <- read.csv("~/directoryLocation/fold0Train.txt")
fold0TreeTrain <- rpart(actualRisk~., method="class",data= fold0TreeTrainData)
summary(fold0TreeTrain)
```

**R code for Binary Tree Classification test:**

```r
fold0TreeTestData  <- read.csv("~/directoryLocation/fold0Test.txt")
fold0TreeTest <- predict(fold0TreeTrain, fold0TreeTestData, type="class")
table(fold0TreeTest, fold0TreeTestData$actualRisk)
```
To plot the Binary Tree:

```r
plot(fold0TreeTrain, uniform=TRUE, branch=0.6, margin=.05)
text(fold0TreeTrain, all=TRUE, use.n=TRUE, cex=.8)
```

**Code to create six parameters scatterplot (as in Figure 5.9):**

**Note:** The csv file should be in following format

```
age,sex,behav,gNode,gPat,gClus,actualRisk
0.627272727,M,Y,0.013738963,0.979840417,1,Diabetic
0.618181818,M,Y,0.024879487,0.671667611,0.523809524,Diabetic
0.663636364,M,N,0.02403873,0.625824694,0,Diabetic
0.427272727,M,Y,0.12641904,0,0,Non-diabetic
0.636363636,M,Y,0,0,0,Non-diabetic
0.7,F,N,0.028718945,0,0,Non-diabetic
0.645454545,F,Y,0.010333195,0,0,Non-diabetic
```

It is then fetched in R as `eachPntScore_forScatterplotOfRFs`

**Actual R code is below**

```r
#utility function for multiple plot
multiplot <- function (...plotlist=NULL, file, cols=1, layout=NULL){
    library(grid)
    plots <- c(list(...),plotlist)
    numPlots = length(plots)
    if(is.null(layout)){
        layout <- matrix(seq(1,cols*ceiling(numPlots/cols)),
                          ncol=cols,nrow = ceiling(numPlots/cols))
    }
    if(numPlots==1){
```

243
print(plots[[1]])

else{
    grid.newpage()
    pushViewport(viewport(layout=grid.layout(nrow(layout),ncol(layout))))
    for(i in 1:numPlots){
        matchidx <- as.data.frame(which(layout==i, arr.ind = TRUE))
        print(plots[[i]],vp = viewport(layout.pos.row=matchidx$row,
                                          layout.pos.col=matchidx$col))
    }
}

#MAIN CODE
library(ggplot2)
alphaVal = 0.5
jitterWidth=0.45
jitterHeight=0.2
effectiveData=eachPntScore_forScatterplotOfRFs  #this is the dataframe
d0=data.frame(values=effectiveData$age,type=effectiveData$actualRisk)
d1=data.frame(values=effectiveData$sex,type=effectiveData$actualRisk)
d2=data.frame(values=effectiveData$behav,type=effectiveData$actualRisk)
d3=data.frame(values=effectiveData$gNode,type=effectiveData$actualRisk)
d4=data.frame(values=effectiveData$gPat,type=effectiveData$actualRisk)
d5=data.frame(values=effectiveData$gClus,type=effectiveData$actualRisk)

g0 = ggplot(d0,aes(type,values))
g0 = g0 + geom_jitter(alpha =
                        alphaVal,aes(colour=type),position=position_jitter(width=jitterWidth))+labs(y="Age",x=element_blank())
g0 = g0 + theme(legend.title=element_blank())
g0 = g0 + theme(legend.position="none")
g1 = ggplot(d1,aes(type,values))
g1 = g1 + geom_jitter(alpha = alphaVal,aes(colour=type),position=position_jitter(width=jitterWidth,height=jitterHeight))+labs(y= "Sex",x=element_blank())
g1 = g1 + theme(legend.title=element_blank())
g1 = g1 + theme(legend.position="none")


g2 = ggplot(d2,aes(type,values))
g2 = g2 + geom_jitter(alpha = alphaVal,aes(colour=type),position=position_jitter(width=jitterWidth,height=jitterHeight))+labs(y= "Behaviour risk",x=element_blank())
g2 = g2 + theme(legend.title=element_blank())
g2 = g2 + theme(legend.position="none")


g3 = ggplot(d3,aes(type,values))
g3 = g3 + geom_jitter(alpha = alphaVal,aes(colour=type),position=position_jitter(width=jitterWidth)) + labs(y= "Node match score",x=element_blank())
g3 = g3 + scale_y_continuous(limits=c(0,0.3)) #we magnified the upper limit of y axis. Very few points are missed in the process.
g3 = g3 + theme(legend.title=element_blank())
g3 = g3 + theme(legend.position="none")


g4 = ggplot(d4,aes(type,values))
g4 = g4 + geom_jitter(alpha = alphaVal,aes(colour=type),position=position_jitter(width=jitterWidth)) + labs(y= "Pattern match score",x=element_blank())
g4 = g4 + theme(legend.title=element_blank())
g4 = g4 + theme(legend.position="none")
g5 = ggplot(d5,aes(type,values))
g5 = g5 + geom_jitter(alpha =
alphaVal,aes(colour=type),position=position_jitter(width=jitterWidth))+labs(y="Cluster match score",x=element_blank())
g5 = g5 + theme(legend.title=element_blank())
g5 = g5 + theme(legend.position="none")

multiplot(g0,g1,g2,g3,g4,g5, cols=2)
Score vs Outcome scatterplot (as in Figure 5.12)

The demo data for the csv file input is like this

score, predictionResult, actualRisk
0.26075246, tp, Diabetic
0.193418201, tp, Diabetic
0.148513129, fn, Diabetic
0.141614969, fn, Diabetic
0.143728813, tn, Non-diabetic
0.1206, tn, Non-diabetic
0.180534846, fp, Non-diabetic

It is fetched in R as scoreVsOutcomeVsPredictionData

Main Code below (y intercept draws the straight line)

library(ggplot2)
d0=data.frame(values=scoreVsOutcomeVsPredictionData$score,type=factor(scoreVsOutcomeVsPredictionData$predictionResult,levels=c("tp", "tn", "fp", "fn")))
g0 = ggplot(d0,aes(type,values))+geom_hline(yintercept=0.1601525,linetype="dashed",colour="red")# draw the threshold line
g0 = g0 + geom_jitter(aes(colour=scoreVsOutcomeVsPredictionData$actualRisk),position=position_jitter(width=0.45))+labs(x="Outcome",y="Score",title="Score vs Outcome")
g0
Appendix I: Accuracy and performance related result for all three predictive models in different settings

**Binary Logistic Regression with Attribution adjustment with Selective Comorbidity**

<table>
<thead>
<tr>
<th></th>
<th>TP</th>
<th>TN</th>
<th>FP</th>
<th>FN</th>
<th>Precision</th>
<th>Recall</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fold 0</td>
<td>193</td>
<td>191</td>
<td>39</td>
<td>37</td>
<td>83.18966</td>
<td>83.91304</td>
<td>83.47826</td>
</tr>
<tr>
<td>Fold 1</td>
<td>190</td>
<td>201</td>
<td>29</td>
<td>40</td>
<td>86.75799</td>
<td>82.6087</td>
<td>85</td>
</tr>
<tr>
<td>Fold 2</td>
<td>188</td>
<td>192</td>
<td>38</td>
<td>42</td>
<td>83.18584</td>
<td>81.73913</td>
<td>82.6087</td>
</tr>
<tr>
<td>Fold 3</td>
<td>185</td>
<td>9</td>
<td>46</td>
<td>45</td>
<td>80.08658</td>
<td>80.43478</td>
<td>80.21739</td>
</tr>
<tr>
<td>Fold 4</td>
<td>182</td>
<td>185</td>
<td>45</td>
<td>48</td>
<td>80.17621</td>
<td>79.13043</td>
<td>79.78261</td>
</tr>
<tr>
<td>Fold 5</td>
<td>189</td>
<td>195</td>
<td>35</td>
<td>41</td>
<td>84.375</td>
<td>82.17391</td>
<td>83.47826</td>
</tr>
<tr>
<td>Fold 6</td>
<td>193</td>
<td>193</td>
<td>37</td>
<td>37</td>
<td>83.91304</td>
<td>83.91304</td>
<td>83.91304</td>
</tr>
<tr>
<td>Fold 7</td>
<td>184</td>
<td>196</td>
<td>34</td>
<td>46</td>
<td>84.40367</td>
<td>80</td>
<td>82.6087</td>
</tr>
<tr>
<td>Fold 8</td>
<td>188</td>
<td>194</td>
<td>36</td>
<td>42</td>
<td>83.92857</td>
<td>81.73913</td>
<td>83.04348</td>
</tr>
<tr>
<td>Fold 9</td>
<td>189</td>
<td>186</td>
<td>44</td>
<td>41</td>
<td>81.11588</td>
<td>82.17391</td>
<td>81.52174</td>
</tr>
<tr>
<td>Mean</td>
<td>188.1</td>
<td>174.2</td>
<td>38.3</td>
<td>41.9</td>
<td>83.11324</td>
<td>81.78261</td>
<td>82.65622</td>
</tr>
<tr>
<td>SD</td>
<td>3.419064</td>
<td>55.24274</td>
<td>5.1</td>
<td>3.419064</td>
<td>1.990811</td>
<td>1.48655</td>
<td>1.548827</td>
</tr>
</tbody>
</table>

**Binary Logistic Regression without Attribution adjustment with Selective Comorbidity**

<table>
<thead>
<tr>
<th></th>
<th>TP</th>
<th>TN</th>
<th>FP</th>
<th>FN</th>
<th>Precision</th>
<th>Recall</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fold 0</td>
<td>175</td>
<td>189</td>
<td>41</td>
<td>55</td>
<td>81.01852</td>
<td>76.08696</td>
<td>79.13043</td>
</tr>
<tr>
<td>Fold 1</td>
<td>182</td>
<td>186</td>
<td>44</td>
<td>48</td>
<td>80.53097</td>
<td>79.13043</td>
<td>80</td>
</tr>
<tr>
<td>Fold 2</td>
<td>174</td>
<td>187</td>
<td>43</td>
<td>56</td>
<td>80.18433</td>
<td>75.65217</td>
<td>78.47826</td>
</tr>
<tr>
<td>Fold 3</td>
<td>168</td>
<td>167</td>
<td>63</td>
<td>62</td>
<td>72.72727</td>
<td>73.04348</td>
<td>72.82609</td>
</tr>
<tr>
<td>Fold 4</td>
<td>167</td>
<td>185</td>
<td>45</td>
<td>63</td>
<td>78.77358</td>
<td>72.6087</td>
<td>76.52174</td>
</tr>
<tr>
<td>Fold 5</td>
<td>172</td>
<td>180</td>
<td>50</td>
<td>58</td>
<td>77.47748</td>
<td>74.78261</td>
<td>76.52174</td>
</tr>
<tr>
<td>Fold 6</td>
<td>175</td>
<td>186</td>
<td>44</td>
<td>55</td>
<td>79.90868</td>
<td>76.08696</td>
<td>78.47826</td>
</tr>
<tr>
<td>Fold 7</td>
<td>166</td>
<td>194</td>
<td>36</td>
<td>64</td>
<td>82.17822</td>
<td>72.17391</td>
<td>78.26087</td>
</tr>
<tr>
<td>Fold 8</td>
<td>179</td>
<td>194</td>
<td>36</td>
<td>51</td>
<td>83.25581</td>
<td>77.82609</td>
<td>81.08696</td>
</tr>
<tr>
<td>Fold 9</td>
<td>180</td>
<td>188</td>
<td>42</td>
<td>50</td>
<td>81.08108</td>
<td>78.26087</td>
<td>80</td>
</tr>
<tr>
<td>Mean</td>
<td>173.8</td>
<td>185.6</td>
<td>44.4</td>
<td>56.2</td>
<td>79.71359</td>
<td>75.56522</td>
<td>78.13043</td>
</tr>
<tr>
<td>SD</td>
<td>5.287722</td>
<td>7.337575</td>
<td>7.337575</td>
<td>5.287722</td>
<td>2.789531</td>
<td>2.299009</td>
<td>2.242821</td>
</tr>
</tbody>
</table>
### Binary Logistic Regression with Attribution adjustment with ICD code collapse

<table>
<thead>
<tr>
<th></th>
<th>TP</th>
<th>TN</th>
<th>FP</th>
<th>FN</th>
<th>Precision</th>
<th>Recall</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fold 0</td>
<td>173</td>
<td>162</td>
<td>68</td>
<td>57</td>
<td>71.78423</td>
<td>75.21739</td>
<td>72.82609</td>
</tr>
<tr>
<td>Fold 1</td>
<td>177</td>
<td>159</td>
<td>71</td>
<td>53</td>
<td>71.37097</td>
<td>76.95652</td>
<td>73.04348</td>
</tr>
<tr>
<td>Fold 2</td>
<td>162</td>
<td>173</td>
<td>57</td>
<td>68</td>
<td>73.9726</td>
<td>70.43478</td>
<td>72.82609</td>
</tr>
<tr>
<td>Fold 3</td>
<td>212</td>
<td>229</td>
<td>1</td>
<td>18</td>
<td>99.53052</td>
<td>92.17391</td>
<td>95.86957</td>
</tr>
<tr>
<td>Fold 4</td>
<td>221</td>
<td>227</td>
<td>3</td>
<td>9</td>
<td>98.66071</td>
<td>96.08696</td>
<td>97.3913</td>
</tr>
<tr>
<td>Fold 5</td>
<td>225</td>
<td>227</td>
<td>3</td>
<td>5</td>
<td>98.68421</td>
<td>97.82609</td>
<td>98.26087</td>
</tr>
<tr>
<td>Fold 6</td>
<td>225</td>
<td>229</td>
<td>1</td>
<td>5</td>
<td>99.55752</td>
<td>97.82609</td>
<td>98.69565</td>
</tr>
<tr>
<td>Fold 7</td>
<td>221</td>
<td>226</td>
<td>4</td>
<td>9</td>
<td>98.22222</td>
<td>96.08696</td>
<td>97.17391</td>
</tr>
<tr>
<td>Fold 8</td>
<td>170</td>
<td>174</td>
<td>56</td>
<td>60</td>
<td>75.22124</td>
<td>73.91304</td>
<td>74.78261</td>
</tr>
<tr>
<td>Mean</td>
<td>195.6</td>
<td>198</td>
<td>32</td>
<td>34.4</td>
<td>86.22255</td>
<td>85.04348</td>
<td>85.56522</td>
</tr>
<tr>
<td>SD</td>
<td>25.66009</td>
<td>29.96998</td>
<td>29.96998</td>
<td>25.66009</td>
<td>12.76742</td>
<td>11.15656</td>
<td>11.95107</td>
</tr>
</tbody>
</table>

### Binary Logistic Regression without Attribution adjustment with ICD code collapse

<table>
<thead>
<tr>
<th></th>
<th>TP</th>
<th>TN</th>
<th>FP</th>
<th>FN</th>
<th>Precision</th>
<th>Recall</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fold 0</td>
<td>172</td>
<td>157</td>
<td>73</td>
<td>58</td>
<td>70.20408</td>
<td>74.78261</td>
<td>71.52174</td>
</tr>
<tr>
<td>Fold 1</td>
<td>179</td>
<td>152</td>
<td>78</td>
<td>51</td>
<td>69.64981</td>
<td>77.82609</td>
<td>71.95652</td>
</tr>
<tr>
<td>Fold 2</td>
<td>171</td>
<td>142</td>
<td>88</td>
<td>59</td>
<td>66.02317</td>
<td>74.34783</td>
<td>68.04348</td>
</tr>
<tr>
<td>Fold 3</td>
<td>166</td>
<td>142</td>
<td>88</td>
<td>64</td>
<td>65.35433</td>
<td>72.17391</td>
<td>66.95652</td>
</tr>
<tr>
<td>Fold 4</td>
<td>167</td>
<td>143</td>
<td>87</td>
<td>63</td>
<td>65.74803</td>
<td>72.6087</td>
<td>67.3913</td>
</tr>
<tr>
<td>Fold 5</td>
<td>169</td>
<td>142</td>
<td>88</td>
<td>61</td>
<td>65.75875</td>
<td>73.47826</td>
<td>67.6087</td>
</tr>
<tr>
<td>Fold 6</td>
<td>171</td>
<td>152</td>
<td>78</td>
<td>59</td>
<td>68.6747</td>
<td>74.34783</td>
<td>70.21739</td>
</tr>
<tr>
<td>Fold 7</td>
<td>166</td>
<td>146</td>
<td>84</td>
<td>64</td>
<td>66.4</td>
<td>72.17391</td>
<td>67.82609</td>
</tr>
<tr>
<td>Fold 8</td>
<td>175</td>
<td>148</td>
<td>82</td>
<td>55</td>
<td>68.09339</td>
<td>76.08696</td>
<td>70.21739</td>
</tr>
<tr>
<td>Fold 9</td>
<td>180</td>
<td>154</td>
<td>76</td>
<td>50</td>
<td>70.3125</td>
<td>78.26087</td>
<td>72.6087</td>
</tr>
<tr>
<td>Mean</td>
<td>171.6</td>
<td>147.8</td>
<td>82.2</td>
<td>58.4</td>
<td>67.62188</td>
<td>74.6087</td>
<td>69.43478</td>
</tr>
<tr>
<td>SD</td>
<td>4.779121</td>
<td>5.344156</td>
<td>5.344156</td>
<td>4.779121</td>
<td>1.884771</td>
<td>2.077879</td>
<td>2.00472</td>
</tr>
</tbody>
</table>
### Parameter Estimation Method with Attribution adjustment with Selective Comorbidity

<table>
<thead>
<tr>
<th>Fold</th>
<th>TP</th>
<th>TN</th>
<th>FP</th>
<th>FN</th>
<th>Precision</th>
<th>Recall</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>210</td>
<td>181</td>
<td>49</td>
<td>20</td>
<td>81.08108</td>
<td>91.30435</td>
<td>85</td>
</tr>
<tr>
<td>1</td>
<td>208</td>
<td>187</td>
<td>43</td>
<td>22</td>
<td>82.86853</td>
<td>90.43478</td>
<td>85.86957</td>
</tr>
<tr>
<td>2</td>
<td>219</td>
<td>180</td>
<td>50</td>
<td>11</td>
<td>81.41264</td>
<td>95.21739</td>
<td>86.73913</td>
</tr>
<tr>
<td>3</td>
<td>217</td>
<td>172</td>
<td>58</td>
<td>13</td>
<td>78.90909</td>
<td>94.34783</td>
<td>84.56522</td>
</tr>
<tr>
<td>4</td>
<td>220</td>
<td>170</td>
<td>60</td>
<td>10</td>
<td>78.57143</td>
<td>95.65217</td>
<td>84.78261</td>
</tr>
<tr>
<td>5</td>
<td>209</td>
<td>187</td>
<td>43</td>
<td>21</td>
<td>82.93651</td>
<td>90.86957</td>
<td>86.08696</td>
</tr>
<tr>
<td>6</td>
<td>208</td>
<td>197</td>
<td>33</td>
<td>22</td>
<td>86.30705</td>
<td>90.43478</td>
<td>88.04348</td>
</tr>
<tr>
<td>7</td>
<td>201</td>
<td>188</td>
<td>42</td>
<td>29</td>
<td>82.71605</td>
<td>87.3913</td>
<td>84.56522</td>
</tr>
<tr>
<td>8</td>
<td>218</td>
<td>177</td>
<td>53</td>
<td>12</td>
<td>80.4428</td>
<td>94.78261</td>
<td>83.86957</td>
</tr>
<tr>
<td>9</td>
<td>204</td>
<td>180</td>
<td>50</td>
<td>26</td>
<td>80.31496</td>
<td>88.69565</td>
<td>83.47826</td>
</tr>
<tr>
<td>Mean</td>
<td>211.4</td>
<td>181.9</td>
<td>48.1</td>
<td>18.6</td>
<td>81.55601</td>
<td>91.91304</td>
<td>85.5</td>
</tr>
<tr>
<td>SD</td>
<td>6.327717</td>
<td>7.673982</td>
<td>7.673982</td>
<td>6.327717</td>
<td>2.156062</td>
<td>2.751181</td>
<td>1.233779</td>
</tr>
</tbody>
</table>

### Parameter Estimation Method without Attribution adjustment with Selective Comorbidity

<table>
<thead>
<tr>
<th>Fold</th>
<th>TP</th>
<th>TN</th>
<th>FP</th>
<th>FN</th>
<th>Precision</th>
<th>Recall</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>659</td>
<td>518</td>
<td>172</td>
<td>31</td>
<td>79.30205</td>
<td>95.50725</td>
<td>83.91304</td>
</tr>
<tr>
<td>1</td>
<td>658</td>
<td>531</td>
<td>159</td>
<td>32</td>
<td>80.53856</td>
<td>95.36232</td>
<td>87.6087</td>
</tr>
<tr>
<td>2</td>
<td>660</td>
<td>525</td>
<td>165</td>
<td>30</td>
<td>80</td>
<td>95.65217</td>
<td>86.95652</td>
</tr>
<tr>
<td>3</td>
<td>655</td>
<td>514</td>
<td>176</td>
<td>35</td>
<td>78.8207</td>
<td>94.92754</td>
<td>83.04348</td>
</tr>
<tr>
<td>4</td>
<td>646</td>
<td>512</td>
<td>178</td>
<td>44</td>
<td>78.39806</td>
<td>93.62319</td>
<td>84.78261</td>
</tr>
<tr>
<td>5</td>
<td>645</td>
<td>528</td>
<td>162</td>
<td>45</td>
<td>79.92565</td>
<td>93.47826</td>
<td>84.3783</td>
</tr>
<tr>
<td>6</td>
<td>674</td>
<td>498</td>
<td>192</td>
<td>16</td>
<td>77.8291</td>
<td>97.68116</td>
<td>86.73913</td>
</tr>
<tr>
<td>7</td>
<td>653</td>
<td>533</td>
<td>157</td>
<td>37</td>
<td>80.61728</td>
<td>94.63768</td>
<td>84.56522</td>
</tr>
<tr>
<td>8</td>
<td>656</td>
<td>520</td>
<td>170</td>
<td>34</td>
<td>79.41889</td>
<td>95.07246</td>
<td>87.17391</td>
</tr>
<tr>
<td>9</td>
<td>649</td>
<td>525</td>
<td>165</td>
<td>41</td>
<td>79.72973</td>
<td>94.05797</td>
<td>83.69565</td>
</tr>
<tr>
<td>Mean</td>
<td>655.5</td>
<td>520.4</td>
<td>169.6</td>
<td>34.5</td>
<td>79.458</td>
<td>95</td>
<td>85.28261</td>
</tr>
<tr>
<td>SD</td>
<td>7.940403</td>
<td>9.951884</td>
<td>9.951884</td>
<td>7.940403</td>
<td>0.854789</td>
<td>1.150783</td>
<td>1.579793</td>
</tr>
</tbody>
</table>
### Parameter Estimation Method with Attribution adjustment with ICD code collapse

<table>
<thead>
<tr>
<th></th>
<th>TP</th>
<th>TN</th>
<th>FP</th>
<th>FN</th>
<th>Precision</th>
<th>Recall</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fold 0</td>
<td>178</td>
<td>168</td>
<td>62</td>
<td>52</td>
<td>74.16667</td>
<td>77.3913</td>
<td>75.21739</td>
</tr>
<tr>
<td>Fold 1</td>
<td>184</td>
<td>159</td>
<td>71</td>
<td>46</td>
<td>72.15686</td>
<td>80</td>
<td>74.56522</td>
</tr>
<tr>
<td>Fold 2</td>
<td>182</td>
<td>154</td>
<td>76</td>
<td>48</td>
<td>70.54264</td>
<td>79.13043</td>
<td>73.04348</td>
</tr>
<tr>
<td>Fold 3</td>
<td>225</td>
<td>227</td>
<td>3</td>
<td>5</td>
<td>98.68421</td>
<td>97.82609</td>
<td>98.26087</td>
</tr>
<tr>
<td>Fold 4</td>
<td>224</td>
<td>227</td>
<td>3</td>
<td>6</td>
<td>98.67841</td>
<td>97.3913</td>
<td>98.04348</td>
</tr>
<tr>
<td>Fold 5</td>
<td>225</td>
<td>227</td>
<td>3</td>
<td>5</td>
<td>98.68421</td>
<td>97.82609</td>
<td>98.26087</td>
</tr>
<tr>
<td>Fold 6</td>
<td>225</td>
<td>227</td>
<td>3</td>
<td>5</td>
<td>98.68421</td>
<td>97.82609</td>
<td>98.26087</td>
</tr>
<tr>
<td>Fold 7</td>
<td>224</td>
<td>225</td>
<td>5</td>
<td>6</td>
<td>97.81659</td>
<td>97.3913</td>
<td>97.6087</td>
</tr>
<tr>
<td>Fold 8</td>
<td>222</td>
<td>223</td>
<td>7</td>
<td>8</td>
<td>96.94323</td>
<td>96.52174</td>
<td>96.73913</td>
</tr>
<tr>
<td>Fold 9</td>
<td>176</td>
<td>173</td>
<td>57</td>
<td>54</td>
<td>75.53648</td>
<td>76.52174</td>
<td>75.86957</td>
</tr>
<tr>
<td>Mean</td>
<td>206.5</td>
<td>201</td>
<td>29</td>
<td>23.5</td>
<td>88.18935</td>
<td>89.78261</td>
<td>88.58696</td>
</tr>
</tbody>
</table>

### Parameter Estimation Method without Attribution adjustment with ICD code collapse

<table>
<thead>
<tr>
<th></th>
<th>TP</th>
<th>TN</th>
<th>FP</th>
<th>FN</th>
<th>Precision</th>
<th>Recall</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fold 0</td>
<td>168</td>
<td>157</td>
<td>73</td>
<td>62</td>
<td>69.70954</td>
<td>73.04348</td>
<td>70.65217</td>
</tr>
<tr>
<td>Fold 1</td>
<td>198</td>
<td>132</td>
<td>98</td>
<td>32</td>
<td>66.89189</td>
<td>86.08696</td>
<td>71.73913</td>
</tr>
<tr>
<td>Fold 2</td>
<td>202</td>
<td>124</td>
<td>106</td>
<td>28</td>
<td>65.58442</td>
<td>87.82609</td>
<td>70.86957</td>
</tr>
<tr>
<td>Fold 3</td>
<td>197</td>
<td>121</td>
<td>109</td>
<td>33</td>
<td>64.37908</td>
<td>85.65217</td>
<td>69.13043</td>
</tr>
<tr>
<td>Fold 4</td>
<td>211</td>
<td>97</td>
<td>133</td>
<td>19</td>
<td>61.33721</td>
<td>91.73913</td>
<td>66.95652</td>
</tr>
<tr>
<td>Fold 5</td>
<td>186</td>
<td>120</td>
<td>110</td>
<td>44</td>
<td>62.83784</td>
<td>80.86957</td>
<td>66.52174</td>
</tr>
<tr>
<td>Fold 6</td>
<td>203</td>
<td>118</td>
<td>112</td>
<td>27</td>
<td>64.44444</td>
<td>88.26087</td>
<td>69.78261</td>
</tr>
<tr>
<td>Fold 7</td>
<td>181</td>
<td>132</td>
<td>98</td>
<td>49</td>
<td>64.87455</td>
<td>78.69565</td>
<td>68.04348</td>
</tr>
<tr>
<td>Fold 8</td>
<td>200</td>
<td>121</td>
<td>109</td>
<td>30</td>
<td>64.72492</td>
<td>86.95652</td>
<td>69.78261</td>
</tr>
<tr>
<td>Fold 9</td>
<td>199</td>
<td>142</td>
<td>88</td>
<td>31</td>
<td>69.33798</td>
<td>86.52174</td>
<td>74.13043</td>
</tr>
<tr>
<td>Mean</td>
<td>194.5</td>
<td>126.4</td>
<td>103.6</td>
<td>35.5</td>
<td>65.41219</td>
<td>84.56522</td>
<td>69.76087</td>
</tr>
<tr>
<td>SD</td>
<td>11.94362</td>
<td>15.10761</td>
<td>15.10761</td>
<td>11.94362</td>
<td>2.491912</td>
<td>5.192877</td>
<td>2.161814</td>
</tr>
</tbody>
</table>
### Binary Tree Classification with Attribution adjustment with Selective Comorbidity

<table>
<thead>
<tr>
<th>Fold</th>
<th>TP</th>
<th>TN</th>
<th>FP</th>
<th>FN</th>
<th>Precision</th>
<th>Recall</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>223</td>
<td>172</td>
<td>58</td>
<td>7</td>
<td>79.35943</td>
<td>96.95652</td>
<td>85.86957</td>
</tr>
<tr>
<td>1</td>
<td>223</td>
<td>172</td>
<td>58</td>
<td>7</td>
<td>79.35943</td>
<td>96.95652</td>
<td>85.86957</td>
</tr>
<tr>
<td>2</td>
<td>208</td>
<td>195</td>
<td>35</td>
<td>22</td>
<td>85.59671</td>
<td>90.43478</td>
<td>87.6087</td>
</tr>
<tr>
<td>3</td>
<td>217</td>
<td>179</td>
<td>51</td>
<td>13</td>
<td>80.97015</td>
<td>94.34783</td>
<td>86.08696</td>
</tr>
<tr>
<td>4</td>
<td>216</td>
<td>175</td>
<td>55</td>
<td>14</td>
<td>79.7048</td>
<td>93.91304</td>
<td>85</td>
</tr>
<tr>
<td>5</td>
<td>207</td>
<td>183</td>
<td>47</td>
<td>23</td>
<td>81.49606</td>
<td>90</td>
<td>84.78261</td>
</tr>
<tr>
<td>6</td>
<td>222</td>
<td>178</td>
<td>52</td>
<td>8</td>
<td>81.0219</td>
<td>96.52174</td>
<td>86.95652</td>
</tr>
<tr>
<td>7</td>
<td>210</td>
<td>193</td>
<td>37</td>
<td>20</td>
<td>85.02024</td>
<td>91.30435</td>
<td>87.6087</td>
</tr>
<tr>
<td>8</td>
<td>219</td>
<td>185</td>
<td>45</td>
<td>11</td>
<td>82.95455</td>
<td>95.21739</td>
<td>87.82609</td>
</tr>
<tr>
<td>Mean</td>
<td>216</td>
<td>180.5</td>
<td>49.5</td>
<td>13.9</td>
<td>81.46041</td>
<td>93.95652</td>
<td>86.21739</td>
</tr>
<tr>
<td>SD</td>
<td>5.7</td>
<td>7.952987</td>
<td>7.952987</td>
<td>5.7</td>
<td>2.232298</td>
<td>2.478261</td>
<td>1.163399</td>
</tr>
</tbody>
</table>

### Binary Tree Classification without Attribution adjustment with Selective Comorbidity

<table>
<thead>
<tr>
<th>Fold</th>
<th>TP</th>
<th>TN</th>
<th>FP</th>
<th>FN</th>
<th>Precision</th>
<th>Recall</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>216</td>
<td>181</td>
<td>49</td>
<td>14</td>
<td>81.50943</td>
<td>93.91304</td>
<td>86.30435</td>
</tr>
<tr>
<td>1</td>
<td>215</td>
<td>197</td>
<td>33</td>
<td>15</td>
<td>86.69355</td>
<td>93.47826</td>
<td>89.56522</td>
</tr>
<tr>
<td>2</td>
<td>215</td>
<td>186</td>
<td>44</td>
<td>15</td>
<td>83.01158</td>
<td>93.47826</td>
<td>87.17391</td>
</tr>
<tr>
<td>3</td>
<td>229</td>
<td>150</td>
<td>80</td>
<td>1</td>
<td>74.11003</td>
<td>99.56522</td>
<td>82.3913</td>
</tr>
<tr>
<td>4</td>
<td>224</td>
<td>168</td>
<td>62</td>
<td>6</td>
<td>78.32168</td>
<td>97.3913</td>
<td>85.21739</td>
</tr>
<tr>
<td>5</td>
<td>218</td>
<td>177</td>
<td>53</td>
<td>12</td>
<td>80.4428</td>
<td>94.78261</td>
<td>85.86957</td>
</tr>
<tr>
<td>6</td>
<td>220</td>
<td>178</td>
<td>52</td>
<td>10</td>
<td>80.88235</td>
<td>95.65217</td>
<td>86.52174</td>
</tr>
<tr>
<td>7</td>
<td>209</td>
<td>182</td>
<td>48</td>
<td>21</td>
<td>81.32296</td>
<td>90.86957</td>
<td>85</td>
</tr>
<tr>
<td>8</td>
<td>213</td>
<td>190</td>
<td>40</td>
<td>17</td>
<td>84.18972</td>
<td>92.6087</td>
<td>87.6087</td>
</tr>
<tr>
<td>9</td>
<td>218</td>
<td>176</td>
<td>54</td>
<td>12</td>
<td>80.14706</td>
<td>94.78261</td>
<td>85.65217</td>
</tr>
<tr>
<td>Mean</td>
<td>217.7</td>
<td>178.5</td>
<td>51.5</td>
<td>12.3</td>
<td>81.06312</td>
<td>94.65217</td>
<td>86.13043</td>
</tr>
<tr>
<td>SD</td>
<td>5.367495</td>
<td>12.16758</td>
<td>12.16758</td>
<td>5.367495</td>
<td>3.204031</td>
<td>2.333693</td>
<td>1.778893</td>
</tr>
</tbody>
</table>
### Binary Tree Classification with Attribution adjustment with ICD code collapse

<table>
<thead>
<tr>
<th></th>
<th>TP</th>
<th>TN</th>
<th>FP</th>
<th>FN</th>
<th>Precision</th>
<th>Recall</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fold 0</td>
<td>163</td>
<td>201</td>
<td>29</td>
<td>67</td>
<td>84.89583</td>
<td>70.86957</td>
<td>79.13043</td>
</tr>
<tr>
<td>Fold 1</td>
<td>196</td>
<td>162</td>
<td>68</td>
<td>34</td>
<td>74.24242</td>
<td>85.21739</td>
<td>77.82609</td>
</tr>
<tr>
<td>Fold 2</td>
<td>186</td>
<td>174</td>
<td>56</td>
<td>44</td>
<td>76.8595</td>
<td>80.86957</td>
<td>78.26087</td>
</tr>
<tr>
<td>Fold 3</td>
<td>225</td>
<td>227</td>
<td>3</td>
<td>5</td>
<td>98.68421</td>
<td>97.82609</td>
<td>98.26087</td>
</tr>
<tr>
<td>Fold 4</td>
<td>224</td>
<td>227</td>
<td>3</td>
<td>6</td>
<td>98.67841</td>
<td>97.3913</td>
<td>98.04348</td>
</tr>
<tr>
<td>Fold 5</td>
<td>225</td>
<td>227</td>
<td>3</td>
<td>5</td>
<td>98.68421</td>
<td>97.82609</td>
<td>98.26087</td>
</tr>
<tr>
<td>Fold 6</td>
<td>224</td>
<td>229</td>
<td>1</td>
<td>6</td>
<td>99.55556</td>
<td>97.3913</td>
<td>98.47826</td>
</tr>
<tr>
<td>Fold 7</td>
<td>228</td>
<td>224</td>
<td>6</td>
<td>2</td>
<td>97.4359</td>
<td>99.13043</td>
<td>98.26087</td>
</tr>
<tr>
<td>Fold 8</td>
<td>227</td>
<td>222</td>
<td>8</td>
<td>3</td>
<td>96.59574</td>
<td>98.69565</td>
<td>97.6087</td>
</tr>
<tr>
<td>Fold 9</td>
<td>199</td>
<td>162</td>
<td>68</td>
<td>31</td>
<td>74.53184</td>
<td>86.52174</td>
<td>78.47826</td>
</tr>
<tr>
<td>Mean</td>
<td>209.7</td>
<td>205.5</td>
<td>24.5</td>
<td>20.3</td>
<td>90.01636</td>
<td>91.17391</td>
<td>90.26087</td>
</tr>
</tbody>
</table>

### Binary Tree Classification without Attribution adjustment with ICD code collapse

<table>
<thead>
<tr>
<th></th>
<th>TP</th>
<th>TN</th>
<th>FP</th>
<th>FN</th>
<th>Precision</th>
<th>Recall</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fold 0</td>
<td>193</td>
<td>155</td>
<td>75</td>
<td>37</td>
<td>72.01493</td>
<td>83.91304</td>
<td>75.65217</td>
</tr>
<tr>
<td>Fold 1</td>
<td>197</td>
<td>148</td>
<td>82</td>
<td>33</td>
<td>70.60932</td>
<td>85.65217</td>
<td>75</td>
</tr>
<tr>
<td>Fold 2</td>
<td>221</td>
<td>119</td>
<td>111</td>
<td>9</td>
<td>66.56627</td>
<td>96.08696</td>
<td>73.91304</td>
</tr>
<tr>
<td>Fold 3</td>
<td>185</td>
<td>169</td>
<td>61</td>
<td>45</td>
<td>75.20325</td>
<td>80.43478</td>
<td>76.95652</td>
</tr>
<tr>
<td>Fold 4</td>
<td>201</td>
<td>138</td>
<td>92</td>
<td>29</td>
<td>68.60068</td>
<td>87.3913</td>
<td>73.69565</td>
</tr>
<tr>
<td>Fold 5</td>
<td>196</td>
<td>144</td>
<td>86</td>
<td>34</td>
<td>69.50355</td>
<td>85.21739</td>
<td>73.91304</td>
</tr>
<tr>
<td>Fold 6</td>
<td>187</td>
<td>167</td>
<td>63</td>
<td>43</td>
<td>74.8</td>
<td>81.30435</td>
<td>76.95652</td>
</tr>
<tr>
<td>Fold 7</td>
<td>184</td>
<td>170</td>
<td>60</td>
<td>46</td>
<td>75.40984</td>
<td>80</td>
<td>76.95652</td>
</tr>
<tr>
<td>Fold 8</td>
<td>192</td>
<td>152</td>
<td>78</td>
<td>38</td>
<td>71.11111</td>
<td>83.47826</td>
<td>74.78261</td>
</tr>
<tr>
<td>Fold 9</td>
<td>181</td>
<td>176</td>
<td>54</td>
<td>49</td>
<td>77.02128</td>
<td>78.69565</td>
<td>77.6087</td>
</tr>
<tr>
<td>Mean</td>
<td>193.7</td>
<td>153.8</td>
<td>76.2</td>
<td>36.3</td>
<td>72.08402</td>
<td>84.21739</td>
<td>75.54348</td>
</tr>
<tr>
<td>SD</td>
<td>10.92749</td>
<td>16.6</td>
<td>16.6</td>
<td>10.92749</td>
<td>3.238949</td>
<td>4.751082</td>
<td>1.409695</td>
</tr>
</tbody>
</table>
Appendix J: Screenshot of the actual Software used to implement the framework
Chronic Disease Risk Prediction Framework

DB User: alchan
DB Pass: ********

Build Inter-B, Analysis UTILS, Analysis phase 1, Analysis phase 2, Analysis phase 3, Analysis March 16

Validate SQL with Postgres: Start, Cancel
Assess Patients: Start, Cancel
Fetch Patient cohort or cluster info: Fetch cohort, Fetch BN cluster, Cancel
Predict Risk: Start, Cancel
Estimate Parameters: Start, Cancel
Start k-fold validation process: Start

Ready
Total patients processed: 240 / 1580

Initializing filter
Successfully initialized extended baseline filter
Creating Extended Baseline Network
Found 1380 patients for processing ExtendedBN. Processing 10 patients at a time.