A Process Mining Approach in Identifying Patient Disease Trajectories using Electronic Health Records

Guntur Prabawa Kusuma

Submitted in accordance with the requirements for the degree of
Doctor of Philosophy

The University of Leeds

School of Computing

August 2022

Intellectual Property and Publication Statement

The candidate confirms that the work submitted is his own, except where work which has formed part of jointly-authored publications has been included. The contribution of the candidate and the other authors to this work has been explicitly indicated below. The candidate confirms that appropriate credit has been given within the thesis where reference has been made to the work of others.
Part of the work in the chapters of this thesis has appeared in jointly-authored publication as listed below:

Chapter 2

1. **Kusuma, G., Hall, M., Gale, C.P., Johnson, O**. 2018. Process mining in Cardiology: A Literature Review. In International Journal Bioscience Biochemistry Bioinformatics Vol.8(4): 226-236 ISSN: 2010-3638 (2018) DOI: 10.17706/ijbbb.2018.8.4.226-236
   The work in this paper was contributed, written, and presented by Kusuma, G. Supervision, feedback, and general guidance were provided by Johnson, O., Gale, C.P., and Hall, M.
2. **Kusuma, G., Kurniati, A., Rojas, E., McInerney, C., Gale, C.P., Johnson, O.** 2021. Process Mining of Disease Trajectories: A Literature Review. In Proceedings of the 31st Medical Informatics Europe (MIE2021) – Volume 281, 29-31 May 2021, ISBN 978-1-64368-184-9 (print) | 978-1-64368-185-6 (online). DOI: 10.3233/SHTI210200

Chapter 3

Kusuma, G., Sykes, S., McInerney, C., Johnson, O. 2020. Process Mining of Disease Trajectories: A Feasibility Study. In Proceedings of the 13th International Joint Conference on Biomedical Engineering Systems and Technologies - Volume 5 HEALTHINF: HEALTHINF, ISBN 978-989-758-398-8, ISSN 2184-4305, pages 705-712. (2020) DOI: 10.5220/0009166607050712
The work in this paper was contributed, written, and presented by Kusuma, G. Supervision, feedback, and general guidance were provided by Johnson, O., McInerney, C., and Sykes, S.

Chapter 4

Kusuma, G, Kurniati, A, McInerney, C, Hall, M, Gale, CP, Johnson, O. 2020. Process Mining of Disease Trajectories in MIMIC-III: A Case Study. In: Lecture Notes in Business Information Processing. 2nd International Conference on Process Mining (ICPM 2020), 04-09 Oct 2020, Virtual conference managed by the University of Padua. Springer Verlag.
The work in this paper was contributed, written, and presented by Kusuma, G. Supervision, feedback, and general guidance were provided by Johnson, O., Gale, C.P., Hall, M., McInerney, C., and Kurniati, A.

Acknowledgements

I have to start by thanking God, who brought me to this journey which I never imagine I could go, who always guides me and carries me when I was fall, who always be beside me and listen to my prayer.

I'm eternally grateful to Mr. Owen Johnson, my Ph.D. supervisor, for their priceless guidance and encouragement throughout this Ph.D. journey, as well as for the opportunities that I never thought I could have had. To Dr Brandon Bennett, for the guidance during the early stages of my journey, I cannot thank you enough. I am also thankful to my sponsor, The Indonesia Endowment Fund for Education (LPDP) – The Ministry of Finance of Republic of Indonesia, for their financial support to fulfil my dream and for unlocking so many opportunities during my journey. To Indonesia, it is now my duty to give back and to contribute.

This research has been carried out by the Cardiovascular Epidemiology Research Group within the Leeds Institute of Cardiovascular and Metabolic Medicine at the University of Leeds. This group includes Professor Chris P. Gale, Dr. Marlous Hall, Dr. Ciarán McInerney, and myself. My own contributions, which are fully and explicitly indicated in the thesis, have been on process mining of disease trajectories, specifically working with healthcare data, NHS Digital's HES-APC and MIMIC-III. The other members of the group have been working on other research projects with different focuses. Their contributions have been built through discussions during my Ph.D. journey. Special thanks to Angelina Kurniati, Amirah Alharbi, Nik Fatinah, Samantha Sykes, and Samantha Crossfield, members of the electronic health records research group led by Mr Owen Johnson. All of you have my deepest thanks for our friendship; all of you made my adaptation to British culture easier than I thought. A very special thanks to Professor Roy Ruddle, who allowed me to use the 6-screen PC, one of the powerful computers at the University of Leeds. To Dr Eric Rojas, for the knowledge, discussions, and laughs. To Dr Thamer Ba-Dhfari, who always listens to me patiently and who never gets bored going out for lunch with me when Angelina was not around.

Special thanks to Pak Kur - Bu Ambar and family, Ibu Merry – Pak Dough and family, Bu Ika – Pak Brahma and family, Mbak Ida – Mas Daeng and family, Mas dr. Hafidz – Mb dr. Maudy and family, Mas Emye – Mbak Kiki and family, Mas Dani – Mbak

Nanda and family, Uni Bona and family, Mas dr. Bekti – Mbak dr. Mas Hari Agung - Mbak Nanda, Mas Imam – Mbak Mita – Aruna, Bu Riama – Pak Paul and family, Pak Gomez – Bu Tere and family, and Bapak Ibu Kos Mas Arif – Teh Cucu and family, all of you have special place in my mind and my heart, thank you for being super supportive to my family.

I also would like to express my gratitude to all the teachers and Miss Smith the Headteacher of Rosebank Primary School, for educating for my children. Also, the teachers and the headteacher of Whingate Primary school who were very welcoming to my daughter, and also the teachers at the Northern Ballet Academy for the ballet lesson for my daughter. To all the staff at the School of Computing, at LIDA, and at the Refectory and LFC who were so generous when providing meals for me.

Special thanks to the PPI Greater Leeds, the INA-Fellowships, and all the Lurah of LPDP around the world, you were the first aid kit when I felt homesick.

I felt so grateful to have a Ph.D wolf-pack Ima, Bintan, Gaby, Erza, Yana, Hizbul, Saad, Taufik “Opik”, Santosa. I will miss another session of playing Heads-Up!

And finally, I would like to thank my family for their support, understanding, great times, and togetherness in this journey. Ima, Gendis, and Panji, you taught me so hard about being a man, a husband, a father, and your best friend. Endless appreciation for my parents, Papa Yoss, and Mama Putu. To my parents-in-law Bapak Priyanto, Ibu Christiati, my sister and brother-in-law Teyonk and OmYan, thank you for the encouragement, support, and the prayer. This thesis is dedicated to all of you!

Abstract

An electronic health record contains valuable evidence of digital footprint that could help clinicians and medical researchers understand how diseases progress over time. Process mining provides various tools to identify and model the disease trajectory, including implementing PM^2, a process mining framework to conduct process mining projects.

This thesis's aims to: (1) develop methods based on process mining to identify disease trajectory models, (2) study the applicability of the method using different EHR, and 3) mine disease trajectory models from actual EHRs. Three studies were conducted to achieve these aims. The first study was a feasibility study of process mining for identifying disease trajectories using a synthetic data set. The second study identified disease trajectory using an actual EHR from a private teaching hospital in Boston, USA. The third study identified disease trajectories from England's population-wide EHR – the NHS England's Hospital Episode Statistic.

This study demonstrates that process mining is useful for mining disease trajectories from electronic health records using standard tools. The method is feasible for producing relatively clinically relevant disease trajectory models from actual electronic health records, which are readily available in any formal health care information systems. This study intends to expand the rich set of techniques and areas of implementation.

Table of Contents

List of Tables

List of Figures

List of Abbreviations

AKI	:	Acute Kidney Injury
AMI	:	Acute Myocardial Infarction
ANST	:	Any non-significant traces
APC	:	Admitted Patient Care
BC	:	Before Christ
BHF	:	British Heart Foundation
BIDMC	:	Beth Israel Deaconess Medical Center
BPA-H	:	Business Process Analysis in Healthcare
CCG	:	Clinical Commissioning Group
CDS	:	Commissioning Data Set
CI	:	Confidence Interval
CM	:	Clinical Modification
COPD	:	Chronic obstructive pulmonary disease
COVID	:	Coronavirus disease
CP-DQF	:	Care Pathway Data Quality Framework
CPRD	:	Clinical Practice Research Datalink
CVD	:	Cardiovascular disease
CVEPI	:	Cardiovascular Episodes
DARS	:	Data Access Request Service
DF	:	Degree of freedom
DFG	:	Directly-follows graph
iDHM	:	interactive Data-aware Heuristics Miner
DISDATE	:	Discharge date
DISDEST	:	Discharge destination
DISMETH	:	Discharge method
DOB	:	Date of birth
DQF	:	Data Quality Framework
DSA	:	Data Sharing Agreement
DSFC	:	Data Sharing Framework Contract
EHR	:	Electronic Health Records
EPIEND	:	Episode end
EPIKEY	:	Episode key
EPIORDER	:	Episode order
EPISTART	:	Episode start
EPISTAT	:	Episode status
EPITYPE	:	Episode type
EPS	:	Electronic Prescriptions Service
ETL	:	Extract Transform Load
FAE	:	Finished Admission Episodes
FCE	:	Finished Consultant Episode
GB	:	Giga byte
GLOW	:	Global Learning Week

GP	:	General Practitioner
HEALTHINF	:	Heatlth Informatics
HES	:	Hospital Episode Statistics
HESID	:	HES Identifier
HIPAA	:	Health Insurance Portability and Accountability Act
HOMEADD	:	Home address
HSCIC	:	Health & Social Care Information Care
ICD	:	International Classification of Disease
ICPM	:	International Conference of Process Mining
IJBBB	:	International Journal of Bioscience, Biochemistry, and Bioinformatics
IM	:	Inductive Miner
IQR	:	Interquartile range
IRC	:	Integrated Research Campus
LCI	:	Lower confidence interval
LHS	:	Learning health system
LIDA	:	Leeds Instituite of Data Analytics
MIE	:	Medical Informatics Europe
MIMIC	:	Medical Information Mart for Intensive Care
MIT	:	Massachusetts Institute of Technology
MPS	:	Master Person Service
MXML	:	Macromedia FleX Markup Language
MYADMIDATE	:	Month Year Admission Date
MYDOB	:	Month Year Date of birth
MYEPIEND	:	Month Year Episode End
MYEPISTART	:	Month Year Episode Start
NCHS	:	National Center for Health Statistics
NHS	:	National Health Service
ODBC	:	Open Database Connectivity
OR	:	Odd Ratio
OS	:	Operating System
PDM	:	Process Diagnostic Method
PM^2	:	Process Mining Project Methodology
PODS	:	Process-Oriented Data Science
RQ	:	Research Question
RR	:	Relative Risk
SD	:	Standard Deviation
SEED	:	Secure Electronic Environment for Data
SEQ	:	Sequence
SNOMED-CT	:	Systematized Nomenclature of Medicine – Clinical Term
SPELBGIN	:	Spell begin
SPELEND	:	Spell end
SQL	:	Structured Query Language

STROBE	:	Strengthening the Reporting of Observational Studies in Epidemiology
SUS	:	Secondary User Service
UK	:	United Kingdom
US	:	United States
USA	:	United States of America
WEKA	:	Waikato Environment for Knowledge Analysis
WHO	:	World Health Organization
XES	:	eXtensible Event Stream
XML	:	eXtended Markup Language

Chapter 1
Introduction

This chapter gives an introduction to the study of this thesis. It is started with an overview (subchapter 1.1) followed by a problem definition (subchapter 1.2). The next section will define this thesis's aim, objective, hypothesis, and research questions (subchapter 1.3). It is followed by a description of the study approach (subchapter 1.4), including the data sources and the data set overview. This chapter concluded with the organisation of this thesis (subchapter 1.5).

1.1 Overview

The occurrence of diseases, depending on the severity of the disease, can disrupt our health, affecting our lifestyle, reducing the quality of life, and increasing the risk of morbidity or even mortality. With the help of clinicians, the patient is assessed to identify what the diseases are through diagnosis and then confirm the occurrence [1]. Upon confirmation, the diagnostic results are then translated into diagnostic codes and recorded in a database for further reference. People then encounter a medical service for help to restore their healthy and prime condition.

In this modern era, recording the diagnoses can be done in an organised way with the aid of a computer-based information system in the hospital. The increasing capacity of storage and computing power enables us to collect a large amount of health care data electronically involving a high level of detail. Considering the collection of the diagnoses and the temporal aspects, a rich array of data can be analysed to understand the association between diseases. The temporal information of the disease occurrence allows us to see the sequence of diseases as trails and to quantify the frequency of one disease followed by another. By recognising this pattern, we can study disease progression even at the scale of a country's population [2-5]. As the size of the data is increasing rapidly, data analysis improvements are needed. One of the relatively new approaches to conducting the study is process mining.

Process mining is an analytical method to discover, monitor and improve the process by analysing the factual data available in the information systems. Process mining uses event data as the input to be analysed for process discovery, process conformance

and process enhancement [6]. Following the concept of an *event* [7], the occurrence of diseases during the human's life course can be considered as significant changes, whether a change from a healthy condition into an ill state or from a sick already condition into suffering with more comorbidities. These events are readily available within electronic health data and have the components required for process mining analysis.

This thesis uses a process mining approach to identify disease trajectories using electronic health records (EHR) with a spectrum of abstraction from lab-simulated, a hospital, and nationwide clinical data registries. Contributions of this study are:

- Developing methods based on process mining to identify disease trajectories over EHR data,
- Assessing the application of process mining in two different EHR data,
- Providing a generic tool to identify disease trajectories of any EHR data.

This study used two data sets to examine the generalisation of the methods applied using a different set of EHR data. The challenge of this study is to pull insights from two domains: computer science and medicine.

1.2 Problem definition

According to a definition by Jensen et al. (2014) [3], disease trajectory is a sequence of diagnoses based on the temporal information being observed in the patients. The disease trajectories were identified using an electronic data set containing the patients' history of diagnoses – the electronic health records.

All patient encounters with the health care provider are recorded for multiple purposes. Most records are used as the base for billing and kept for future reference for the patients and the clinicians. The records store the diagnoses using standard coding, e.g. the International Classification of Disease (ICD) by the WHO [8]. Each record has temporal information to mark when the disease occurred. The above information is the input for the proposed method in this study to identify the disease trajectories using electronic health records.

Limited attempts have been made to model the disease trajectories using a data-driven approach, natural language processing, and machine learning. There is an opportunity to identify disease trajectories using a process mining method. Process mining is useful for identifying processes, modelling the process, evaluating the model's

quality, and enhancing the operations. More details of these approaches are presented in Chapter 2.

Process mining is an emerging approach to analyse business processes, including discovering process models, conformance checking and process enhancement [9]. Process mining provides a holistic view and end-to-end analysis, generating easy-to-read models and simulations [9]. The input of process mining is an event log, a collection of events created from the readily available EHR. Each event contains at least a case, an activity, and a timestamp. Process mining produces a sequential model as a result of the discovery process. The disease trajectories' sequence of disease has similar properties to the event log. This study's hypothesis is by considering the recorded diagnostic codes in the EHR as the same as the activity in process mining. A range of process mining toolsets is useful for identifying disease trajectories using the EHR.

This thesis analyses the process mining approach for the disease trajectory and then evaluates the quality of the disease trajectory model. The aim is to build and examine a method to analyse the process discovery and process conformance to generate a well-representative disease trajectory model.

1.3 Aim, objective, hypothesis, and research questions

The aim of this study is to explore the feasibility of process mining to identify the disease trajectories using EHR. The objectives to achieve this aim were:

1. To determine the applicability of process mining to discover disease trajectories.
2. To apply process mining methodology as a guideline to plan and execute process mining project to increase reproducibility.
3. To discover disease trajectory models using an actual hospital's electronic health record.
4. To evaluate the robustness of the process mining method in identifying disease trajectory models.
5. To evaluate the representativeness of the process mining method in identifying disease trajectory models.
6. To evaluate the scalability of the process mining method in identifying the disease trajectory model.

7. To use process mining method to identify disease trajectory models of patients diagnosed with acute myocardial infarction.

Towards the aim and the objectives above, the key hypothesis is that *process mining is useful for identifying disease trajectories from routinely collected electronic health records*..

The main research question of this study is, *'Can disease trajectories be identified using a process mining approach?'* (RQ-1). Following this research question, further questions emerged regarding the pattern of the trajectories: *'What are the most followed trajectories and how long was the duration of the trajectories?'* (RQ-2), *'What are the longest and shortest duration time transition trajectories?'* (RQ-3) and *'Are there differences in trajectories followed by different patient groups?'* (RQ-4). A trajectory should be unidirectional and not contain a loop. This research needs to find a method to remove any repeated events and decide the trajectory's direction (RQ-5). Also, the produced disease trajectory model should be representative enough to the data (RQ-6). The essential part of this research is to use EHR as the input, whereas the database was not designed for process mining.

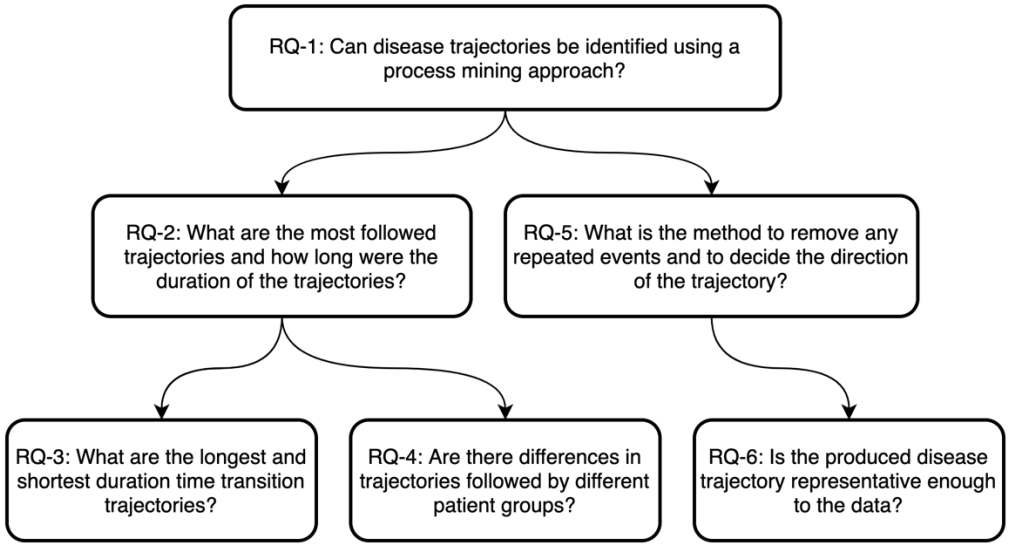


Figure 1.1 The definition of research question.

1.4 Study approach

The general approach of this study is to understand the different formats of electronic health records to identify the components required for mining disease trajectories. These include extracting the data needed, selecting the different abstractions to

represent the trajectories, analysing the association and the directionality between diagnosis codes, and developing the disease trajectories using process mining, including measuring the model's quality.

This current work follows the methodology of the PM^2 framework [10] to conduct disease trajectory mining using process mining. A detailed discussion of this methodology is available in Chapter 3. The PM^2 framework provides six detailed stages of process mining, including each stage's activities. Also, it allows iterations to conduct analyses where we can use each iteration to answer more specific research questions. This work's main research question is, "Can disease trajectories be identified using a process mining approach?"

Following the PM^2 framework, this study started the current work with planning to determine research questions, select a business process for analyses, and compose a team. Subsequently, data Extraction, Transformation, and Load (ETL) were done using available data for this study. Data extraction was done to get the event log from the data source available for this study. This current work did data transformation to transform event logs into pairs of diagnostic codes to enable two statistical analyses: to get the pairs with the significant association and the pairs with trajectories defined. Further transformation is required to transform the pairs of diagnostic codes into the event log for the next step. The selected event log is loaded into process mining tools for disease trajectory identification, process mining analyses, and evaluation.

The event log transformations and the process mining analyses are the main importance of this study. After the event log was transformed into a pair log, this study measured the strength of association between the pairs. Statistical analyses passed only pairs with significantly strong associations into the next steps. The strongly associated pairs may contain pairs that have bidirectional connections. The selected pairs, including pairs with bidirectional relationships, were then analysed using the binomial test to determine any statistically significant direction to define the trajectories. The evaluation with a clinician and an epidemiologist was done by discussion to see if the disease trajectories built using data conform with the evidence or the expert knowledge.

1.4.1 Data sources

This study used three data sets from three different sources of various size and coverage. The first data set is recreated from the subset of Jensen et al.'s (2014)

disease trajectory model; the second data set is from the Beth Israel Deaconess Medical Center (BIDMC) in Boston, USA [11]; and the third data set is from England's National Health Service (NHS) Digital, respectively.

The second data source was the BIDMC – a large non-profit teaching hospital of Harvard Medical School. The hospital is a result of a merger activity in 1996 between the New England Deaconess Hospital (founded in 1896) and the Beth Israel Hospital (founded in 1916). The hospital is located in Boston, covering a land area of 48.34 square miles (125.20 km2) with an estimated population of 692,600 citizens by 1 July 2019 [12]. BIDMC continuously provides patient care as well as educational and research activities. One important mark of BIDMC's contribution to research is that the MIMIC-III health database is freely available for various research studies.

The third data source was the National Health Service (NHS) Digital is the national health care provider in England, specifically those involved with the NHS of England. England spans a land area of 50,301 square miles (130,279 km2) with a population estimate of 56,287,000 citizens based on the mid-year estimate released on 24 June 2020 [13]. England has a relatively similar area size compared to Boston, but its residents are 80 times that of Boston. The NHS Digital runs Spine service, a secure IT infrastructure for patient data in NHS England. The Spine service offers benefits to the patient, including the e-Prescription Service, the Summary Care Records, the e-Referral Service, and the Child Protection – Information Sharing System. The NHS Digital also collects the national "Hospital Episode Statistics" (HES) that records patient care episodes when admitted to the hospital.

1.4.2 Overview of the data sets

The data set being used in this thesis are:

1. **MIMIC-III data set [14]**

   The Medical Information Mart for Intensive Care III (MIMIC-III) data set is available critical care data from the Beth Israel Deaconess Medical Center (BIDMC) hospital in Boston, USA. It covers patient records in the critical units of the BIDMC between 2001 and 2012. The data set contains patient demographics, vital sign measurements, laboratory test results, procedures, and diagnoses. The MIMIC-III data set combines archives from critical care information systems of the BIDMC, hospital EHR databases, and the Social Security Administration Death Master file. This data set has been deidentified and curated by the Laboratory for Computational Physiology team at the

Massachusetts Institute of Technology. This study used the MIMIC-III v1.4 released on 2 September 2016 and will be called the "MIMIC-III data set" from this point forward.

The MIMIC-III data set covers 53,423 distinct hospital admissions for adult patients aged 16 years or above who were admitted to critical care units during 2001-2012 and 7,870 neonates admitted during 2001-2008. The data set consists of several classes: billing, descriptive (demographic), dictionary, interventions, laboratory, medications, notes, physiologic, and reports. One important part of the data for this study is the diagnoses_icd table which contains the diagnostics codes of patients. The diagnoses_icd table is related to the admissions table, which makes it possible to track patients' diagnoses during their admission times. The diagnostic codes are available in ICD-9 format.

2. HES-APC data set

The Hospital Episode Statistics – Admitted Patient Care (HES-APC) data set is a population-wide summary of hospitalisation data in England. Each HES record contains patient information in the NHS hospitals, including their clinical information about diagnoses and procedures; patient information such as dates and methods of admission and discharge; and geographical information such as treatment location and resident area. The published HES data has been deidentified and applied with a statistical disclosure control based on the NHS Digital protocol.

The HES-APC data files are structured based on the financial years. A row in HES-APC represents a Finished Consultant Episode (FCE), a continuous period of care done by a consultant during a specified start and end date. Hospital admissions were from March 2020 to February 2021, HES-APC covers 16 million FCEs, 55.7% (8.9 million) of which included at least one procedure or intervention, and 4.7 million of which were day cases; 12.7 million finished admission episodes (FAEs), of which 5.4 million were emergency admissions. A 'spell' in HES-APC represents a hospital admission, an uninterrupted inpatient stay at one hospital. A spell may include one or more FCEs, depending on the number of consultants visiting during the same stay. This study works on the sequence of diagnoses during patient episodes.

1.5 Conclusion

The current work aimed to provide a method to identify disease trajectories that can be applied using a patient's EHR despite the type of the health care provider or the patient's specific conditions. This study used two sources of patient EHR with different characteristics to achieve the aim. A survey of relevant literature is required to provide a background of this current work for the following methods and findings. The next chapter presents a detailed literature review with the objectives:

1. To determine if process mining conceptually workable to identify disease trajectories.
2. To determine the appropriate methodology to identify disease trajectories.
3. To determine the quality measurement of the disease trajectory model.

1.6 Organisation of the thesis

The structure and topics of each chapter in this thesis are highlighted as follows:

1. Chapter-1: Introduction

   The first chapter gives an overview of the study, followed by a problem definition, aim, objective, hypothesis, and research questions. It also describes the study approaches, including the data sources and overview of the data sets, and concludes with this thesis organisation.

2. Chapter-2: Background

   The second chapter presents the health care as well as the technical backgrounds of the study. The health care background includes health care systems, EHR research, the International Classification of Diseases (ICDs), and disease trajectories. The technical background reviews process mining as the main approach, how it has been applied in health care, process mining of disease trajectories, and the statistical method of this thesis.

3. Chapter-3: Methodology

   Chapter 3 presents the methodology followed in this study. It started with a description of the general methodology, followed by the feasibility study and the advanced stages. The next section describes the overview of the two case studies analysed in this thesis: the MIMIC-III and the HES-APC data sets. The general methodology followed in this study is the Process Mining Project Methodology (PM^2). A description of the adjustment of the general methodology in the two case studies of this thesis will be explained in the

respective case studies. Those two case studies are described in more detail in Chapters 4 and 5. Chapter 3 is closed with a summary of the methodology.

4. Chapter-4: Case study-1, experiments using the MIMIC-III data set

   This chapter describes the first case study using the MIMIC-III data set. The experiment in this case study followed the general methodology, as described in Chapter 3. The description of the experiment in this chapter is structured following the method. The results described in this chapter have been presented at a virtual International Conference of Process Mining (ICPM) in 2020.

5. Chapter-5: Case study-2, experiments using the HES-APC data set

   This chapter explores the second case study covering three experiments using the HES-APC data set. This chapter started with data description, including the data provenance, characterisation, quality, and representativeness of the data set. It is followed by the description of three experiments in process mining of disease trajectories: an experiment using a sample of five thousand randomly selected patients, an experiment using a stratified random sample of five thousand patients, and an experiment on acute myocardial infarction patients. The exploration of each experiment is presented following the stages in the general methodology.

6. Chapter-6: Discussion

   This chapter presents a discussion based on the previous chapters of this thesis. The discussion includes a reflection on the method, the answers to research questions, the challenges of using health care data, disease trajectory mining, and the impact of the disease. Some challenges of using health care data described in this study are data access and confidentiality, data quality, data understanding, and data and disease trajectory visualisation.

7. Chapter-7: Summary

   The final chapter of this thesis gives some conclusions, presentations and feedback, future work, and the final remark on this study. Conclusions are structured following the general steps in this study, including the literature review, methodology development, the experiment on the MIMIC-III data set, the experiment on the HES-APC data set, and the discussion.

Chapter 2
Background

In Chapter 2, the background of this Ph.D. study is presented to cover both perspectives of health care and technical aspects. This chapter includes two publications of literature reviews: first publication is “Process mining in cardiology: a literature review” and the second is “Process mining of disease trajectories: a literature review” which are summarised in sections 2.2.2.3 and 2.2.3 respectively. The first publication was presented at the International Conference on Information Technology 2017 where the latter was selected for a publication in a journal, the International Journal of Bioscience, Biochemistry, and Bioinformatics (IJBBB). The second publication was presented at the 31st Medical Informatics Europe 2021 (MIE 2021) virtual conference as summarised in section 2.2.3.

2.1 Health care background

The study of identifying human disease trajectory is gaining interest in recent years and since then many approaches have been proposed to identify disease trajectories. A recent approach is mapping disease proteins into a graph to identify disease trajectories, known as *interactomics* [15]. The protein in this context is a functional molecule produced by a gene. The interactions between proteins develop a complex cellular mechanism to perform many functions in the human body. If these interactions collapse, then it will have an effect on human health. This recently proposed method is fulfilling one of the Hills’s criteria of causality [16] and opens more opportunities to unlock the complicated area of causality. This section covers health care systems, electronic health records, standard coding for classifying diseases, and the disease trajectories research.

2.1.1 Health care systems

Health care is an essential part of human to allow them pursue life goals, get help if they are in pain or suffering, prevent the early deaths, and be the source of information for better future planning [17]. A health care system, according to the World Health Organisation (WHO), comprises of organisations, people and actions with the main intention is to promote, restore or maintain health [18]. The system is aimed to

improve health level and equity, responsiveness, social and financial risk protection, and resource usage efficiency.

Looking at how health care has developed over the past years, Hood et al. promoted a hypothetical vision of predictive, preventive, personalised, and participatory ('P4') [19] as a substitution for a health care that is reactive to disease. This new vision emerged based on three developments: system biology and system medicine, digital revolution, and consumer-driven health care and social network. The development of technology enables the health care system to improve the quality of care whilst effectively reducing costs, decreasing disease incidence, and advancing the health care system to establish a learning health system (LHS) [20] as defined by the Institute of Medicine (IoM) [21].

2.1.2 Electronic Health Record research

One element of the health care system is data that made available in the form of electronic health records (EHR). Using the current technology, the source of data for EHR is increasing in volume and computational power. This condition makes it possible to use EHR as a source for medical informatics researchers as its secondary use.

One of the earliest and most famous medical records has been found in the "Edwin Smith" papyrus. The papyrus is dated back to 1600 BC but is believed to be a copy of the original record dated from 3000 BC [22]. It contains medical records where mostly about the surgery of wounds of the human's upper body parts [23]. The early purpose of having a record of diseases is to journal the doctor's knowledge when they found a way to cure the diseases so the journal can be used as a reference and passed through generations.

How the patient's medical records are being used and manipulated is changing. Jump into the computer era, where many healthcare organisations rely their business process on computers, and data can be saved in the form of bits "1" and "0". Many countries are supported by the power of health information systems to collect, store, analyse, and then use the outcome to support the clinician's decisions. The patients could receive high-quality care services using the health information system. At the same time, the hospital management could also benefit from the system to help them make a high-quality decisions to help their operational activities, including planning the

hospital's development. Further, the system could help the government create the rules and policies to shape the country to be better and ready for future challenges.

The emerging computer technology allows the combination and collaboration of the medical and computer science domains. Using the technology, we can now have a massive amount of medical record data which can be seen as a valuable source when data science methods come into play. Hemmingway et al. [24] identified the potential and challenges of big data research on cardiovascular disease. One of the challenges identified is to discover models of disease networks.

2.1.3 International Classification of Diseases

The International Classification of Diseases (ICD) is a classification standard of diagnostic codes that has been used worldwide for clinical and research purposes. ICD is the foundation for the identification of health trends as well as the statistical analysis of diseases and mortality. This standard is maintained by the World Health Organisation (WHO) as an authority for health within the United Nations System [8]. The ICD was designed as a health care classification system to map health conditions based on the main categories and specific variations. The first version of the ICD was introduced in 1893 in France. It has been changed and improved several times. The major version- the ICD-9- was intended to update the classification with little changes to avoid expenses on data adaptations (1978). It was followed by a Clinical Modification (ICD-9-CM) created by the US National Center for Health Statistics (NCHS) to assign diagnostic and procedure codes. The ICD-10 was the next version endorsed by the Forty-third World Health Assembly (1994) allowing more codes and tracking many new diagnoses and procedures than those in the ICD-9. The ICD-11 is the latest version released in June 2018, that comes with an implementation package including transition tables from and to ICD-10, a coding tool, a manual, and more supports [25].

In this study, the MIMIC-III data used the ICD-9 codes while the HES data used the ICD-10 codes. Understanding the ICD-9 and the history of ICD in general is important to identify possible issues that may arise in the analysis of this thesis.

2.1.4 Disease trajectories

Disease trajectories describe the time-related relations between diseases or diagnoses and their progressions. This topic has become an interest in analysing health processes and is currently an emergent topic in the literature. A systematic review by Pinaire et al. (2021) explored 70 articles published between 2000 and 2015 in the trajectory concept. The awareness of this important topic evidenced in the studies in America, Asia, Europe, and intercontinental Australasia. Those studies analysed a range of trajectory concepts, including cost, care, care process, health outcomes, biological measure, risk, and survival. Some diseases commonly analysed in those studies are cardiovascular diseases, cancer, diabetes, neurological diseases, lung diseases, and kidney diseases [26].

In 2014, Jensen et al. [3] reported a study of temporal disease trajectories in a large-scale national hospital database of Danish population. The database analysed in this study consists of 6.2 million patients recorded for 14.9 years. This study focused on providing health recommendations based on the health outcomes recorded in the registry database. Their study identified 1,171 trajectories, created pathology-centred clusters using ICD-10, and computed Related Risks (RR) to measure diagnosis pair association. From the strongly associated pairs, a binomial test was used to determine the trajectory.

Following the above, Jensen et al. constructed a disease trajectory model by overlapping pairs of diagnostic codes to make a longer trajectory. For example, two pairs of trajectory A → B and B → C were overlapped to construct a trajectory of A→ B → C. The result of this "overlapping" approach wasn't checked if such trajectory is actually existed.

A feasibility study by Kusuma et al. in 2020 [27] showed process mining as a potential approach to support disease trajectory studies. Based on an event log of disease extracted from an EHR data, the sequence of the first occurrence of diseases of each patient can be studied. This feasibility study was a background study supporting this thesis.

2.2 Technical background

Process mining is an emerging approach to enable process exploration including to find actual processes or process enhancement. A range of tools and techniques in

process mining is available for implementation using four different perspectives [6]: control-flow, performance, conformance, and organisational perspectives.

There are three types of exploration in process mining: process discovery to produce process models from event log data, process conformance to confront process models to event log or vice versa, and process enhancement where information of the actual process recorded in the event log is used to improve the process model.

2.2.1 Process mining

Process mining is an emerging approach coined by Will van der Aalst for analysing processes. These processes are mined to discover, monitor, and improve by extracting knowledge from the already available event log within an information system [28]. The event log contains the actual processes that had happened in the organisation as a result of following guidelines or new processes as part of innovations.

There are three types of process mining available: discovery, conformance checking, and enhancement. The three process mining types will be elaborated in the next sub-section. The goals of process mining are to identify any previously unknown processes, to control the trailing process in the system, or to quantify the processes that defy the pre-designed guidelines [29]. Furthermore, process mining has several mining perspectives when it comes into implementation, those are [9]:

1. The control-flow perspective focuses on the ordering of activities to find the good characteristics of all possible paths.
2. The organisational perspective focuses on the resources being involved in the processes to classify the people and their respected roles including the social network.
3. The case perspective focuses on the properties of the case where a case can be characterised by its path in the process, or by the people working on it.
4. The time perspective focuses on the timing and the frequency of events. It is possible to identify bottlenecks, measure the service levels, monitor the utilisation or resources, and predict the remaining processing time of a running case.

The focus of this research are the control-flow and the time perspectives to understand the patterns and the duration of the disease trajectory models.

2.2.1.1 Process discovery

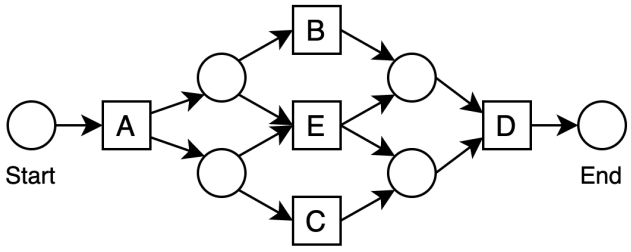
Process discovery refers to the combination of the discovery task and the control-flow perspective. Process discovery is the most challenging task where a process model is constructed by capturing the behaviour seen in the event log. The challenge is to find the process discovery algorithm that can map the event log into a process model so that the model is representative of the event log. Below is the description of four examples of process discovery algorithms to show the various approach to discover process model:

1) Alpha (α) miner

The alpha miner algorithm is also known as α-algorithm [30] that receives an event log as input and returns an output of a Place/Transition net (P/T-net). The idea behind this algorithm is to examine the observed causal relationships between tasks. There are four types of relationships between the tasks and their respective notations: direct succession or follow (>), causality (→), parallel (||), and choice (#). A footprint matrix is composed based on the tasks and the relationship to construct a process model. Figure 2.1 shows the example of the footprint matrix and the constructed process model using the α-algorithm with ingoing and outgoing arcs.

	A	B	C	D	E
A	#	→	→	#	→
B	←	#	||	→	#
C	←	||	#	→	#
D	#	←	←	#	←
E	←	#	#	→	#

(a)



(b)

Figure 2.1 (a) A footprint matrix and (b) a process model constructed using the α-algorithm.

The α-algorithm is the first process discovery algorithm and is able to handle concurrency. This algorithm's benefit is the ability to be workable on a structured process. The limitations are: it couldn't correctly mine incomplete event logs, noisy event logs, unstructured processes, and short loop or duplicate tasks [31].

2) Fuzzy miner

The Fuzzy miner algorithm [32] was designed and proposed to answer the limitation of the Alpha miner which is handling unstructured processes. The idea behind this algorithm is by adopting the concept of roadmap where the level of details can be adjusted. The resulting process model is represented by nodes (rectangles) and connected by arrows. The rectangle represents the *significance* to measure the relative importance of activities, while the arrows represent the *association* to measure the relation of two activities (behaviour) following one another.

The significance and association can be defined based on the level of interest in details. For the simplified model, only the highly significant activities and relations are preserved while the less significant but highly correlated relations are aggregated.

3) Inductive miner

The Inductive Miner (IM) algorithm is designed based on *process trees* [33]. A process tree is a hierarchical representation of a block-structured workflow net that describes a language. The tree has *leaves* which labelled with *activities* and it also has *nodes* which labelled with as *operators* describing how the subtrees are combined. The Inductive miner has a set of operators ($\oplus$) containing: $\times$ means the exclusive choice between one of the subtrees, $\rightarrow$ means the sequential execution of the subtrees, $\circlearrowleft$ means the structured loop, and $\wedge$ means a parallel execution.

IM discovers process models using a divide and conquer approach by splitting the event log to get sub-logs recursively. An operator from $\oplus$ is defined every time a split has been made. These sub-logs, if combined again using $\oplus$ will produce the initial event log. Using this approach, IM can handle existing infrequent behaviours. This algorithm produces process models with the characteristics of high fitness and high generalisation but low in precision.

4) Heuristics miner

The Heuristics Miner algorithm [34] includes the analysis of causal dependencies and a causal matrix to produce a process model. The algorithm has three steps to discover process model: (1) mining of the dependency graph from event log, (2) determine the input and out expression of each activity based on the type of dependencies, and (3) mining long distance dependencies.

The heuristics miner is one of the algorithms that offers a good performance [35] and is able to deal with noise and low frequent behaviour, or only the main behaviour. The limitation is that a heuristics miner assumes the event log is complete and doesn't contain a noise event log (noise-free).

2.2.1.2 Conformance checking

The Conformance checking is a process mining task to evaluate the resulting process model with regard to the event log in term of conformity [6]. Process model, as produced by discovery activity using algorithms, still needs to be evaluated. Non-conformances between process model and event log could occur in two ways: (1) when the discovered process model does not reflect the actual behaviours, or (2) when some behaviours found in the event log deviate from the model.

The quality of the discovered process model can be evaluated using four quality dimensions: fitness, precision, generalisation, and simplicity [9, 36]. Each dimension has a range of value from 0 to 1 where 1 represents a good value. A good process model has high values in all four dimensions.

Conformance checking is relatively similar to machine learning in measuring the model's quality. Both techniques require a model that is not underfitting or overfitting. The difference between process mining and machine learning is in the way of learning the data. Process mining does not identify other patterns unless the sequence of event data. The measures in conformance checking are basically by calculating the extent to which a series of activities can be retraced or "played" in the model. The cost of non-replayable activity in the event log to the model or vice versa is calculated. In machine learning, data is represented as a value in a multi-dimensional space and then classified using statistical methods to build a model for prediction [37]. Machine learning can be used in conjunction with process mining to build a model for prediction, e.g. what activity comes next or to predict if a processing time will exceed the predefined limit [38]. Despite their differences, process mining and machine learning produce model where the quality can be measured.

The measurement of a process model in process mining is by using conformance checking, while measurement of a model in machine learning is by building a confusion matrix. The following describes each of the quality dimensions of conformance checking [39]:

1. Fitness

The fitness dimension (Q_{rf}) measures the extent to which the discovered process model can replay the behaviour seen in the event log. A process model will get the fitness value of 1 if all traces in the event log can be replayed by the model.

The technique to compute fitness is by aligning as many events as possible from the trace with the activities when the model is executed. Events in the log may be skipped, or activities may be inserted without any corresponding event in the event log. This skipping and insertion will cost penalties to the overall score. The following is the formula to compute fitness:

$$Q_{rf} = 1 - \frac{cost\ for\ aligning\ model\ and\ event\ log}{minimal\ cost\ to\ align\ arbitrary\ event\ log\ on\ model\ or\ vice\ versa}$$

The denominator is the minimal costs if there is no match exists between event log and process model.

2. Precision

The precision dimension measures how a model does not allow too much behaviour to be seen in the event log and is not underfitting. Imprecision happens when the model plays more behaviours than those suggested in the event log. An underfitting model happens if a model allows for behaviours very different from what was seen in the event log. Precision score 1 means that behaviours produced by the model are available in the event log. The precision score (Qp) [40] is calculated as follows:

$$Q_p = \frac{the\ number\ of\ observed\ activities\ in\ the\ context}{total\ number\ of\ activities\ possible\ in\ the\ model}$$

where the context as the numerator is related to the level of precision being measured in the log level or precision being measured in the case level.

Other metrics for precision calculation are varied, the soundness metric [41], the behavioural appropriateness metric [42], and the alignment-based precision metric [40, 43]. The alignment-based precision metric no longer estimates precision based on the model, but on an aligned event log to the model. It compares the number of different occurrences of activities to the total possible number of activities observed in the model.

Precision in machine learning is the probability of data predicted as True Positive being classified correctly.

3. Generalisation

In general, a process model should not restrict behaviour that only seen in the event log [40]. This type of model is called "overfitting", which means that it only fits to the examples in the event log. The measurement of the generalisation comes with various metrics: the alignment-based probabilistic, the frequency of use, and the behavioural generalisation. The alignment-based probabilistic is related to the Alignment-based Fitness and Alignment-based Precision, it estimates the probability, using the Bayesian statistics, that a new unobserved case can be replayed by the existing model.

4. Simplicity

The Simplicity dimension is based on the fact that simpler models are preferred rather than the complex one. There are two interpretations of "simpler" model: 1) models that are not extremely large and the density of arcs is low [44], and 2) model as understandability that put more emphasis on the ease of interpretation and cognitive capabilities. The effort to simplify a process model is mainly related to the pre-processing steps e.g., by filtering the number of activities taken in the process discovery.

2.2.1.3 Enhancement

Enhancement is the third type of process mining that extends or improves the discovered process model. Here the information from different perspectives that already included in the event log are added. Three known perspectives were incorporated for process enhancement: the organisational perspective, time perspective, and case perspective [45].

The practice of process enhancement began with conformance checking [6]. If the process model does not represent the actual behaviour seen in the event log, then model repairment or model extension is required. Yasmin et al. [45] identified that the trend of process repair mostly happened in the control-flow perspective while the organisation, time and case perspectives were used for process enhancement. The

most studied domain for process enhancement practice is medical followed by governmental and financial studies.

2.2.1.4 Process mining methodology

Process mining methodology helps researchers or practitioners to conduct process mining projects. After two decades of process mining development, there are various process mining methodologies that provide guidelines to support reproducibility of the process mining works. The first process mining methodology is the L* life-cycle model [46], which has been improved to include iteration analysis. The improved methodology is the Process Mining Project Methodology (PM^2) [10]. The PM^2 is then extended to become the ClearPath method [47] by including a process simulation for better engagement with the expertise. Another methodology is the Process Diagnostic Method (PDM) [48], which could provide an extensive outline of the process within the information system quickly. An extension also happened to the PDM and became the Business Process Analysis in Health care (BPA-H) for mining processes in health care environment [49]. The Question-Driven methodology [50] is another method of conducting process mining in the health care environment where the analysis is led by questions.

Among the methodologies above, three have been selected as the main methodologies: the L* Life-cycle model, the PM^2, and the Question-driven methodology. The L* Life-cycle model was chosen since it provides the basic steps of conducting a process mining project in a sequential way. The PM^2 was chosen since it allows iteration and provides detailed steps for data processing. Finally, the Question-driven methodology was chosen since it provides the frequently-posed questions from medical experts that are useful for constructing research questions.

This research identifies disease trajectories where only the first-ever occurrence will be considered. The domain-specific knowledge is valuable to understand how the diseases have progressed. This research requires collaboration with the domain expert. Therefore, the PDM was not chosen because the method only considers the event log as the source of information. Any prior knowledge including the knowledge that is specific to a specialised domain is dismissed. The BPA-H was also excluded since it requires sequence clustering analysis for sequences with repeated events, while disease trajectory only considers the first disease occurrence as the event. The three selected methodologies are described below.

1) L* Life-cycle model

The L* life-cycle model is the first introduced methodology to conduct process mining projects [6, 46]. This model consists of five stages, the first stage is the *Planning and justification* (Stage 0) to initiate the project by understanding the data and the business. The second stage (Stage 1) is to *extract* the event data from the system, any predefined models if available, and other directives. Those inputs are needed for the third stage (Stage 2) *creating a control-flow model and connecting the event log*. The process model from the Stage 2 is enhanced in the fourth stage (Stage 3) to create an integrated process model by adding more perspectives. The fifth stage (Stage 4) used the Stage 3 results for *operational support* activities: detect, predict, and recommend.

The L* life-cycle model is mainly to discover a single process and then enrich it with performance and resource information. Thus, the L* life-cycle model is suitable for structured processes [10]. Another limitation is that the L* life-cycle model does not offer iterations.

2) Process Mining Project Methodology

The Process Mining Project Methodology (PM^2) [10] consists of six stages as seen in illustrated in Figure 2.2. The PM^2 is suitable for both structured and unstructured processes [51]. It consists of six stages where five of them are divided into two groups: the initialisation (Stages 1 and 2), and the analysis iterations (Stages 3 until 5).

The *Planning* (Stage 1) consists of determining the research question, selecting business processes, and forming the project team. Determining scope is done in the *Extraction* (Stage 2), including extracting event data, and transferring process knowledge. The *Data Processing* (Stage 3) is to create event logs that are optimal for the mining and analysis stage by creating views, aggregating events, enriching event logs, and filtering logs. The *Mining and Analysis* (Stage 4) consists of process discovery, conformance checking, enhancement, and process analytics. The *Evaluation* (Stage 5) is to diagnose, verify and validate the analysis findings to improve ideas that fulfilled the project goals. The results from the evaluation stage, such as the improvement ideas, become the input to Stage 6 *Process Improvement and Support* stage (Stage 6). The final stage consists of two activities: implementing improvements and supporting operations to achieve the output: process modifications. Figure 2.2 presents the overview of the PM^2 methodology.

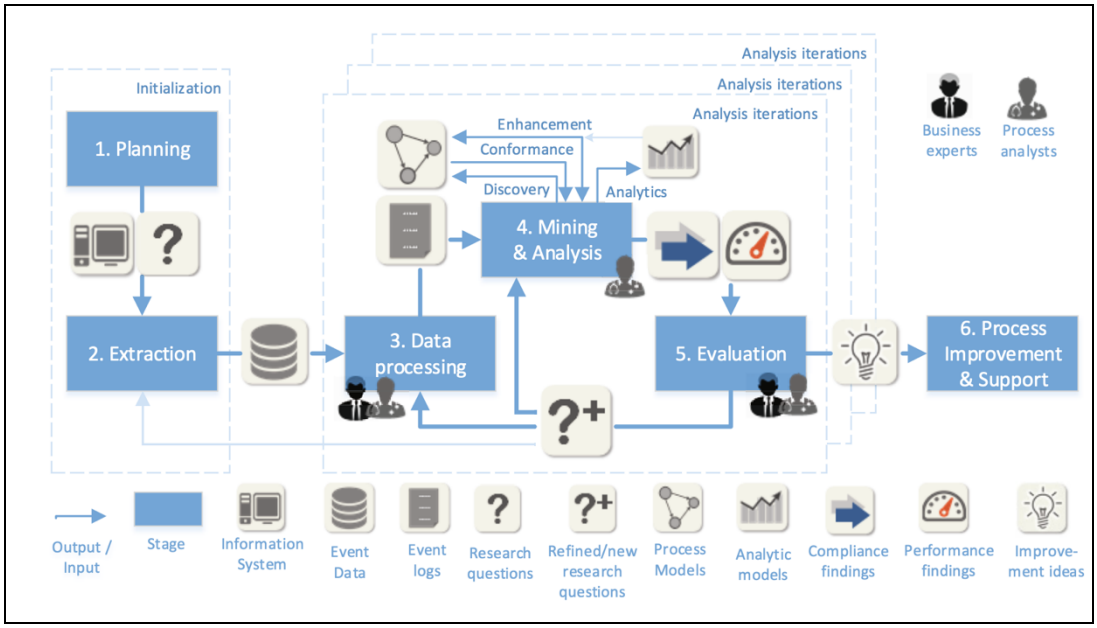


Figure 2.2 The overview of PM^2 methodology reproduced from [10]

The PM^2 methodology was designed to be iterative to improve process performance or compliance with rules and regulations. The need for collaboration between process analysts and business experts is also suggested. The PM^2 provides detailed directions in every stage that are useful and easy-to-follow in real projects. The PM^2 methodology is adopted in this study. Adjustments since in Stage 1 are required to analyse disease trajectory instead of processes in organisations to achieve the specific purpose of the experiments.

3) Question-driven methodology

The Question-driven methodology is proposed by Rojas et al. [50] and designed for conducting process mining in the health care environment. This method is started with questions that are frequently posed by experts in the domain regarding the process execution. The questions are responded using a process mining approach to find the answers. This methodology contains six stages and each of the stages is described below.

Stage 1 Data extraction contains activities of data identification, data extraction from the Hospital Information System, ensuring all required fields are available in the data including timestamps, and a verification of the data quality. *Stage 2 Event log creation* this stage includes the identification of the frequently posed questions from the domain experts to create an event log. The created event log requires further

inspection if the desired characteristics are available and with the correct value. *Stage 3 Filtering*, this stage refines the event log by filtering that covers the basic filtering, clinical filtering, and question-driven filtering. Stage 4 *Data analysis* includes the activities of selecting the data analysis techniques, using the selected statistical analysis to characterise the event log, and data mining analyses. *Stage 5 Process mining* includes identifying the tool, process discovery, conformance checking, performance analysis, organisational analysis, analysis on the type of the questions, and data analysis and process mining cycle. *Stage 6 Results evaluation* identifies domain experts who are responsible for the analysis, defines feedback instruments, and obtains the feedback.

2.2.1.5 Process mining tools

This current work demonstrates the feasibility of using process mining tools to identify disease trajectories. It is important to show that disease trajectory mining can be done using a standard process mining tool. DISCO was chosen since it is lightweight and easy to acquire an academic license for on-premise installation to keep the research data safe. The DISCO supports ease-of-use and easy installation that could help explain the disease trajectories. Celonis is another process mining tool that provides ease-of-use. Compared to DISCO, Celonis has more capabilities to do conformance checking and create a bespoke dashboard for interactive use. It is also available in the form of online tools. DISCO and Celonis are commercial tools, while an open-source process mining tool of ProM Framework is also available. The ProM Framework was chosen since it is a standard tool used by the researchers worldwide to develop new process mining techniques (https://www.promtools.org). As a framework, ProM Framework provides various plugins to support the required types of analyses. Scientific documentation of every plugin in ProM is mostly available on the internet including the online ProM Forum (https://www.win.tue.nl/promforum) to allow interaction with the community of process miners.

1) ProM framework

ProM Toolkit is a process mining tool mainly for academic or research purposes [52, 53]. ProM is a popular open-source process mining framework where the algorithms and other features are available as plugins. Plugins are modules containing the implementation of process mining techniques (including discovery algorithms,

conformance checking, etc.) or visualisations which developed by members of the community around the world. ProM receives event logs in various formats: the eXtensible Event Stream (XES) [54], comma-separated values (CSV), or the generic eXtensible Markup Language (XML) format.

The ProM framework is available for computers running Windows, macOS, or other Linux based platforms. This tool comes with two separate software. The first software is the ProM itself and the second is the ProM Package Manager for installing or updating objects or packages and managing memory allocation to support the execution of any ProM's functionalities.

2) DISCO

DISCO is a commercial process mining software produced by Fluxicon [55]. It is a tool that offers a quick and easy way to produce process models including the process statistics, and to perform event log filtering. DISCO relies on one process discovery algorithm to produce process models, the DISCO Miner algorithm based on the Fuzzy miner algorithm [32]. This tool receives event logs in standard formats XES, CSV or MXML and produces process models that can be viewed by frequency or duration including animation. DISCO is also able to import and export event log compatible to other process mining tools such as ProM framework.

3) Celonis

Celonis is also a commercial process mining software which was founded in 2011 as a startup company [56]. The Munich-based software company create Celonis to map and analyse business processes to identify bottlenecks, inefficiencies, and opportunities for improvement. This tool uses an event log as the input data from the company's systems, including enterprise resource planning (ERP), customer relationship management (CRM), or as a file of various formats such as CSV, Excel, or XML.

2.2.2 Process mining in health care

2.2.2.1 Process mining in general health care domain

An initial literature study of process mining in health care was carried out by Yang and Su in 2014 [57]. They identified 37 articles of how process mining has been used to improve the quality of clinical pathways during the period of 2004 to 2013. The 37

articles were classified based on three categories: process discovery, variants analysis and control, and evaluation and improvement. Yang and Su identified four key trends from their study: further analysis of the variants, integrated process management, customisation, and self-learning improvement of the clinical pathway.

Rojas et al. in 2015 [58] conducted a systematic overview of different approaches that have been used to analyse health care processes using process mining. The number of articles was not directly stated but there were 30 articles have been covered. They analysed the type of data, the frequently-posed questions, the methods or algorithms, the used methodologies, medical fields, and geographical location. In 2016, Rojas et al. [59] extended the work by reviewing 74 articles including the articles from the previous work and also extending the analyses from six into eleven aspects. The added aspects are process types, process mining perspectives, process mining tools, implementation strategies, and analysis strategies. To date, the extended version of Rojas et al. has become the most cited process mining in health care literature study.

Following the work by Rojas et al., two literature reviews of process mining in health care were produced in 2016. First, a systematised literature review by Ghasemi and Amyot, where they provided an overview of process mining in general and an example of application in the health care environment. Ghasemi and Amyot also provided three insights: first, the number of publications in the domain has been growing; second, among the obtained articles of literature review, three publications were conducted in a systematic way; and third, the challenge to produce a good quality of literature reviews should provide a comprehensive and up to date view, conducted in a systematic fashion and should explicitly covers the validity. The second literature review is by Erdoğan and Tarhan [60] who produced a systematic study on the studies that used conformance checking technique and provide an overview of how their developed analytic software have been validated in industrial contexts.

Erdoğan and Tarhan had extended their work in 2018 [61] by conducting systematic mapping study using 172 articles published during 2005 and 2017. They created a concept map based on the following attributes: types of research and contribution, application context, health care speciality, mining activity, process modelling type and notion/ language, and mining algorithm. The result shows that the domain is rapidly growing, and more studies reported validations of their respective works. On the other hand, Erdoğan and Tarhan had suggested a challenge of a lacking study of process mining in terms of multiple departments in a hospital or multiple hospitals.

Another literature study in 2018 was done by Batista and Solanas [62] who decided to explain the current trends from the 55 examined articles rather than to explain the articles in detail. They classified the trends into nine dimensions covering the objective of the process mining analysis, the process mining type, perspective, algorithms & tools, medical facilities, medical fields, medical process type, medical data, and medical data pre-processing techniques. Finally, the most recent literature study was from Dallagassa et al. in 2021 where a systematic mapping study was conducted. They studied 270 articles that were published between 2002 and 2019. They provided the evolution of process mining in health since 2001 until 2019 including the classification based on the health care environment, main areas of application, strategy/ algorithm adopted, and the main contribution of process mining application to the health care field where the most dominant is on the performance evaluation, bottleneck, and time management and scheduling.

2.2.2.2 Process mining in specific health care domain

Process mining is also found for implementation in specific health care domains. These specific implementation covers the studies of Mannhardt and Blinde [63], and Hendricks [64] whose implemented process mining in a similar setting of the emergency room for patients with Sepsis. Kurniati et al. [65] used process mining to analyse cancer pathways; Farid et al. [66] studied literature review of the process mining implementation in the domain of frail elderly care; Williams et al. [67] conducted a literature review of process mining implementation based on a health care setting which is primary care; while the other specific domain were reported in a literature study by Helm et al. [68] that covers all chapters of the ICD-10.

2.2.2.3 Process mining in cardiology

As in other health care domains, process mining has also been applied in cardiology. Cardiology is the study to diagnose and treat disorders of heart and blood vessels, including coronary heart disease, rheumatic heart disease, and other cardiac conditions [69]. In the previous work in 2018 [70], there were 32 papers carefully selected and reviewed as previous studies related to process mining in cardiology. The 32 selected papers were published between 2008 and 2017. Those papers have been

reviewed based on five thematic analyses: (1) the process and data types; (2) research questions; (3) process mining perspectives, types, and tools; (4) methodologies; and (5) limitations and future work. The following paragraph summarises the results of that literature review paper, and the final paragraph concludes this section with insights into this thesis.

In term of process types, there were seven out of 32 papers [71] applied in process mining in organisational processes. There were 24 out of 32 papers on medical treatment process, and one paper [72] applied process mining to both organisational and medical treatment processes. There were 20 out of 32 papers analysed data from more than the one hospital for comparative analysis [73]. Out of those 32 papers, the most common CVD diagnoses analysed are stroke (9 papers) and unstable angina (9 papers). The research questions of those previous studies were dominated by the proposal of new approaches, as were found in 15 out of 32 selected papers [74]. Other papers applied an algorithm or a combination of algorithms to solve a specific problem in various case studies. Our analysis on process mining perspective, types and tools found that there were sixteen studies discussed control-flow perspective, fourteen studies discussed the control-flow and performance perspective., and only one study [72] discussed all three types, which are control-flow, performance, and organisational perspective. Seven studies [73] used ProM toolkit, with two of them combined it with another tools including WEKA, or Rapid Miner. Despite the wide range of process mining implementation in cardiology, we found that there was a low awareness of process mining methodologies, as evidenced that there was only one paper [73] reported a methodology followed in their paper (the L* life-cycle methodology). The limitation and further work derived from those selected studies were related to data and technique. Data limitations included noises, missing data, and data quality control. Technical limitations were limitations in computer processing power, memory usage, bias associated with the mining algorithm, and simplified control flow due to computational complexity. Suggestions for future work were to improve data with additional health parameters, analysis of different CVD diagnoses, and additional data from other departments; to improve techniques or algorithms; and to improve engagement with domain experts.

Those literature review papers provided evidence of the potential benefits of process mining in cardiology. A multi-disciplinary collaboration with experts is important to make sure that process mining supports medical experts to develop a better

understanding of the actual care pathways, improving quality of the interventions and outcomes. The review highlights the low data quality and low awareness of process mining methodologies as major issues to be addressed.

2.2.2.4 Challenges and opportunities

Process mining in health care is gaining popularity as a field of study. One of the indicators of this trend is the increasing number of published articles in the field [57-59, 75]. Health care processes by nature are complex due to heterogeneity, multidisciplinary, ad-hoc and dynamic to changes [76]. It is even more complex since the health care data containing both structured and unstructured data [24]. The structured EHR data are recorded using standard coding of health conditions, such as the ICD-9, ICD-010, the recently introduced standard ICD-11, or the Systematized Nomenclature of Medicine – Clinical Term (SNOMED-CT). The unstructured part of EHR, for example is the patient medical history in the form of free-text notes, handover notes, imaging reports, etc. In the perspective of process mining, the process model of a complex process is known as a "spaghetti model" [6].

Process mining offers a range of approaches to reducing complexity. Two approaches have been introduced by [77] using *data filtering* or *abstraction*. In the data filtering approach, there are three types of filtering: *event filter* to remove or to keep one or several events from the event log; *event pair filter* to remove or to keep events that fulfil a specific condition to allow a relationship between events; and *trace filter* where a sequence of events that met a defined criteria can be removed or kept. The second approach, abstraction, is by removing the subset of the nodes of a process model to produce a smaller dependency graph. This approach aggregates the paths or activities to provide a better understanding about how the process works on a higher level of abstraction. Another approach to reduce complexity is trace clustering [78]. The idea behind trace clustering is to group the traces in the event log where traces with more similarity will create a "cluster" while the difference between clusters should remain as distinct as possible [79].

The next challenge is the data quality of the EHR for process mining analysis [80]. The EHR is relevant to the understanding of the readily available data within the information system of any formal health care organisation including hospitals. The primary use of EHR is for supporting patient treatment and administrative purposes. The quality of the EHR could also be affected by changes that had happened in the

organisation [81], thus requiring data pre-processing and cleaning for the purpose of research. For this challenge, Weiskopf and Weng had studied various methods of data quality assessment of EHR for research purposes and identified five data quality dimensions [82]: completeness, correctness, concordance, plausibility, and currency. Based on this method, a data quality framework to assess the quality of EHR for process mining analysis has been proposed [80]. Another approach to measure the quality of the data is by using a rating system from one to five stars representing poor quality to excellent quality as described in the Process Mining Manifesto [46]. The highest level of quality is described as trustworthy, complete, well-defined, recorded in an automatic, systematic, reliable, and safe manner, and privacy and security considerations are adequately addressed.

Finally, as described in the highest level of data quality, data privacy is another challenge in the field of health care process mining. Event data often contain highly sensitive or contain private information and this may cause discrimination or other potential harm [24]. Responding to this challenge, an ethical approval is required to work with highly sensitive data [83]. In more technical ways, an event log can be secured based on the abstraction that allows individual traces of a process to remain anonymous but at the same time, a process model and a social network are still discoverable [84]. Another approach is by providing new infrastructures to publish the privacy-aware event data which includes anonymisation operations. Further, a privacy extension for the XES [85] is proposed.

2.2.3 Process mining of disease trajectories

Previous research identifying disease trajectories using a process mining approach was identified and the summary has been published [MIE2021]. This section contains the summary of the paper. The literature review was conducted on 5th November 2020 and a similar approach as [59] was followed. There were 156 potential articles identified from Google Scholar, PubMed, and dblp databases. Additionally, the process mining website at http://processmining.org was also searched for related articles.

A set of filtering steps were carried out, covering the exclusion of duplicates and non-English articles, selection based on titles, abstract, and based on a full reading of the articles. Final selection consist of four process mining articles [27, 86-88] and there were seven features being identified covering: data source, size of the data, selection

criteria, method to identify trajectories, process mining algorithms, model visualisations, and the methods of evaluation.

Each article showcased different sources of data including a variation of the abstraction level. With the abstraction level of an intensive care unit of a private hospital in Boston, USA, Kusuma et al. [86] identified disease trajectories from the freely available MIMIC-III data set [14] using a process mining approach. The data set contains detailed admissions of 46,520 unique patients during their stay in the hospital's intensive care unit. At a higher level of abstraction, de Toledo et al. identified disease trajectories using electronic health care records covering a suburban area in Spain [87]. Their data source came from public health care providers including hospitals, primary cares, and emergency centres containing records of 225,000 patients. De Oliveira et al. managed to identify disease trajectories from electronic health records with the level of abstraction as high as England population using NHS England's data set: Hospital Episode Statistics (HES) [88]. There were 76,523 patients who's electronic health care records being analysed. One more article used a synthetic data set of 50 patients for a feasibility study to see if process mining can be used to identify disease trajectories [27]. The synthetic data were recreated from a subset of a disease trajectory model by Jensen et al. [3].

To identify the disease trajectories, three methods were identified. The first method by Kusuma et al. [27, 86] involved transformation of event log into pairs of diagnostic codes – named "pair log", and then by using the pair log, statistical analyses were conducted to measure the association between the paired diagnostic codes using Relative Risk and then using only strong associated pairs, an identification of trajectory direction is conducted using the binomial test. The second method is n-grams where the trajectories were decided based on the frequency of each gram [87]. The third method is using an iterative model discovery, the proprietary Metaheuristics optimisation algorithm to obtain the disease trajectory model [88].

Analysis of the disease trajectory of patient known with a certain condition was reported in two articles. In [87], disease trajectories were identified from patients with Type-2 Diabetes, while [88] from patients with sepsis. One article of feasibility study used synthetic data to reflect a subset of a disease trajectory of Jensen et al. [3, 27]. In contrast with [87] and [88], Kusuma et al. did not specify certain conditions of patients but used the available data agnostically in identifying the disease trajectories [86].

Process mining discovery algorithms were used to discover disease trajectory models. The most commonly used discovery algorithms are based on the Heuristics approach: Heuristics miner algorithm in [87] and interactive Data-aware Heuristics Miner (iDHM) in [27, 86]. Other algorithms were Fuzzy miner used by [87] and a proprietary Metaheuristics optimisation algorithm used by [88]. As for the visualisation of the models, two articles used Directly-follows graph [27, 86], one article used both Heuristics net and Fuzzy Model [87], and one article used a model visualisation from an app developed by a private company [88].

In term of evaluation, conformance checking was reported in two articles [27, 86] where replay fitness, precision, and generalisation were applied. Further evaluation using k-fold cross-validation was applied in [86], more about cross-validation is provided in the following section the statistical approach. The Metaheuristics optimisation algorithm used in [88] is embedded with 'replayability' measurement, or replay fitness, during the discovery process to get the final best model. In brief, Table 2-1 presents the key similarities and differences between the three methods.

Table 2-1 Key similarities and differences of [27, 86-88].

	Kusuma et al. [27, 86]	de Toledo et al. [87]	De Oliveira et al. [88]
Using EHR	Yes	Yes	Yes
N	46,520	225,000	76,523
Country	USA	Spain	England
Disease selection	No	Yes, Type-2 Diabetes	Yes, Sepsis
Disease trajectory identification	Process mining, event log transformation into pair log, and include statistical analysis	N-grams, clustering process mining	Process mining
Discovery algorithm	iDHM	Heuristics Miner, Fuzzy miner	Metaheuristics optimisation algorithm
Conformance Checking	Yes	No	Yes
Result validation	Cross-validation	Expert consultation	No

2.2.4 Statistical approach

Like other process mining projects that use a statistical approach [48, 89, 90] this thesis also implements some statistical approaches for data processing. Descriptive and inferential statistics are useful approaches in describing event logs, checking the model's conformance to the event log, or comparing two event logs. This current work uses descriptive statistics to describe the collected sample data and uses the inferential statistics for hypotheses testing.

The statistical tests are used along with the other filtering approaches in process mining to remove or keep events in the event log. The filtering activities are required to create the disease trajectories where the statistical approaches include the Pareto principle, and two hypotheses testing consist of Relative Risk and binomial test. These statistical approaches are described in the following sections.

2.2.4.1 Pareto principle

The Pareto principle supports the selection of the commonly occurred diagnostic codes in the data [91, 92]. Due to the large amount of data and the limitation of the computing power to process the data, it is fair to say that only commonly occurred data are taken into account. The Pareto principle uses 80/20 rule to select the commonly occurring data. The rule indicates that 80 percent of problems may be came from as little as 20% of causes. By adopting this principle into this current work, it allows the selection of the 80% of the total incidents of disease to select as little as 20% of the total number of records of diagnostic codes in the data set.

2.2.4.2 Measurement of association

A measure of association determines the relationship between two groups: the exposed group and another group representing the expected level (outcome), by comparison analysis [93]. Relative Risk (RR) is a common measure in medical and epidemiological research that is defined as the probability of some health event in the exposed group divided by the probability of the outcome group with the risk among another group. The definition of a health event can be disease occurrence and the groups are often named as a *case group* versus a *control group*.

The value of RR indicates the likelihood of the risk where the values are described as follows:

1. RR = 1 indicates identical risk among the two groups, or, the risk can not be determined,
2. RR > 1 indicates an increased risk or the risk is likely to happen, and
3. RR < 1 indicates a decreased risk or the risk is unlikely to happen.

Another measure of association, the Odd Ratio (OR), is often used to determine risk and may end up confusing with the RR [94]. The OR compares the probabilities of some events in an exposed group versus the probabilities in a non-exposed group and is calculated as the number of events divided by the number of non-events. The OR returns a less accurate ratio compared to RR. When the two calculations are conducted on rare disease, both OR and RR may be comparable and it is hard to say which one is more correct than the other; on the other hand, if the case happen in more common diseases, then the OR will overestimate the risk. In this case, the RR will be more accurate to estimate risk and OR should be avoided. Additionally, RR is widely used and accepted within the medical community.

The measurement of association described in this section has been implemented in section 3.1.3.3.

2.2.4.3 Pearson Chi-square test

Hypothesis testing is needed to determine the statistical significance of the RR results. Most of the time, research projects managed to work with sample data since it is very difficult to obtain data on the whole population and then measure the likelihood of the risk. The statistically significant test is applied to infer if the association represented by the RR is dominant by random or not even from the sample data.

The Karl Pearson Chi-square (also known as “Chi-square test”) is designed to test the *categorical data* where the data are counted and divided by a certain category [95]. The symbol of the Chi-square test is χ^2 [96]. There are three types of tests that can be done using Chi-square test: (1) test of the independence of variables, (2) test of homogeneity, and (3) test of goodness of fit. The benefit of using the Chi-square test is that statisticians can utilise a statistical method and interpret the findings without relying on the normal distribution. The Chi-square’s significance value was obtained using the degree of freedom and degree of significance, and consulting a Chi-square table [96].

In this current work, Chi-square test is used to test the hypothesis that there is no association between two groups (*null hypothesis*). The threshold of the significance

level (α) is set at 0.05 (5%) following the commonly used threshold. Thus, if the Chi-square significance value is less than the α meaning that the association is less likely to happen by chance, and strong enough to reject the null hypothesis.

2.2.4.4 Binomial test

A Binomial test is a hypothesis test to compare the observed frequencies of the two categories of variable to the expected frequencies under a binomial distribution with a specified probability parameter [97]. The binomial distribution models experiments where a repeated binary outcome is counted [98]. Each binary outcome is called a *Bernoulli* trial. The outcome of the experiments is binary since there are only two possible outcomes e.g., either success or failure, head or tail (for a coin tossing experiment), etc. The probability of the outcome is set as 0.5 (50%) if the chance is divided equally for each outcome.

The Binomial test resulting a probability value (p-value) to determine if the outcome happened purely by chance or otherwise in a systematic fashion. The p-value is then compared using a similar assumption of significance level at 0.05 ($\alpha = 0.05$). The null hypothesis for binomial testing is that the outcomes of the experiments are at random (non-significant, p-value > α). If the p-value is less than the α, then the null hypothesis can be rejected. The implementation of the Binomial test in this current work is provided in section 3.1.3.3.

2.2.4.5 Cross-validation

Cross-validation is a statistical technique for evaluating the generalisability of a model and avoiding overfitting [99, 100]. The goal is to produce a model that is well-suited for a given data set and that can generalise well to unseen data. The technique is often found in the study of machine learning, specifically in supervised learning, where the accuracy of the produced model becomes the main concern [101]. This cross-validation technique is useful when validation using independent study becomes expensive and impractical. The model's predictive ability could be estimated using cross-validation by resampling the data.

In general, cross-validation requires a new set of unseen data (validation data) to evaluate the model's generalisability. This new set of data is taken from the same population as the data for building the model (training data) [102]. The model's

generalisability is quantified by estimation error. A high estimation error indicates that the model needs some adjustment.

There are three types of cross-validation in general:

1. Random sub-sampling

   The random sub-sampling type is known as Monte Carlo cross-validation [103]. The population data is split into multiple random subsets and the data in each split is randomly selected as test data. The rest of the data in each split are used to build a model and then use the test data for assessing the predictive accuracy [104]. The results from each split are averaged.

   The disadvantage of this type of cross-validation is that there will be some data that may never have the chance to be the validation data. The results will be different every time the cross-validation is repeated.

2. Leave-p-out

   The leave-p-out creates p cases as the validation test and the rest as training data [105]. If N is the number of the population data and $p = 1$, meaning that one case becomes the validation data, then the number of the training data is N-1. The number of p for the validation data can vary and the validation process is repeated until all cases are validated. This type of cross-validation is exhaustive since the training and testing should be conducted as many as $\binom{N}{p}$ times.

3. k-fold [105]

   The population data is partitioned into k subsets where the size of each subset is equal. For each fold, a model is built using k-1 subsets as the training data, and the rest would be the validation data. The fold is repeated until each subset becomes the validation data exactly once. The advantage of k-fold cross-validation is that each case is used as validation and training data.

A Process model is a result of process discovery where algorithms are used. Each algorithm may have a different approach to discovering the process model. Therefore, validation is needed to achieve an optimal process model. In process mining, cross-validation can be used to validate the discovered process model from the event log. In contrast to machine learning, the process model validation is conducted by conformance checking, where the measure of replay-fitness, precision, and generalisation are used.

Chapter 3
Methodology

Chapter 2 presented the literature background that supports this study. In this chapter, the methodology to discover disease trajectories is elaborated. Section 3.1 describes the adopted main methodology in this current work that consists of five stages. Section 3.2 describes the feasibility study of using process mining to identify disease trajectories. Section 3.3 describes the data sets for case studies that were carried in this study. Section 3.4 summarises the content of this chapter.

The methodology in this chapter has been presented and published at a conference: the 13th International Conference on Health Informatics. The jointly authored publication entitled "Process Mining of Disease Trajectories: A Feasibility Study".

3.1 PM² for disease trajectory mining

The steps to identify disease trajectories using process mining were mainly based on the Process Mining Project Methodology (PM²) [10]. The PM² framework is designed for conducting process mining projects with multiple analysis iterations. These iterations are extensions to the *L**-Life cycle model [46]. The following are descriptions of each stage of PM², covering the objective, input, output, and activities. Figure 3.1 shows the outline of the PM² framework where the blue rectangles illustrate the stages of the PM² framework, the arrows representing input/ output, and it includes two actors of business experts and process analysts.

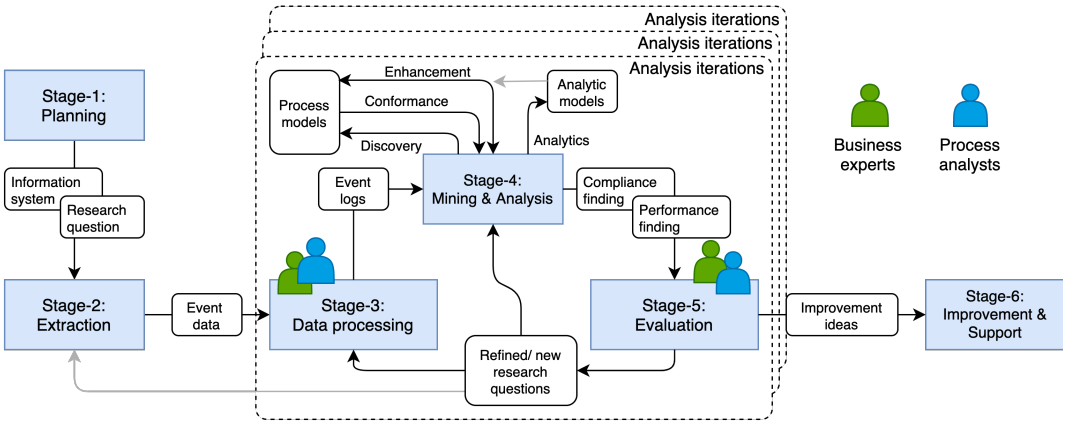


Figure 3.1 The outline of the PM² Framework (recreated from [10]).

3.1.1 Stage-1: Planning

The objective of this stage is to evaluate whether process mining methods are workable to identify disease trajectory models. Three activities conducted in this stage are: selection of business process for analysis, defining research questions, and composing a team.

3.1.1.1 Selection of the scope of process for analysis

The current work analysed disease trajectories instead of business processes. Disease trajectories have the properties for process mining analysis i.e., data refer to cases, events, and temporal information. Based on those properties, the diagnostic codes in electronic health records (EHR) are treated as the equivalent of activities in process mining. The hypothesis, therefore, disease trajectories can be mined using the rich tool set of process mining.

The main characteristic of a disease trajectory is the unidirectionality of the sequence of diseases. Relapsing or reoccurring diseases are excluded, only the first occurrences of each disease are considered. Process mining requires an event log as the input where it contains the history of patients being admitted to hospitals including the main reason for hospitalisation. A quality check of the data is needed to identify if the data have enough information to achieve the goal. Disease trajectory models are discovered to understand how diseases are progressed over time.

Using the information above, patients' EHR is needed and can be retrieved from health care providers' information system. The plan is to create an event log for process mining to contain event data from patients' EHR and to incorporate a method of identifying disease trajectories by Jensen et al. [3] as part of the analysis. The method requires a pairwise analysis where the sequence of diseases is transformed into pairs of diagnostic codes. A formal definition of a disease trajectory model, an event log, and a pair log is provided below:

> **Definition 1 (Disease trajectory model)**. A disease trajectory model ***M*** is a directed acyclic graph to model the trace ***T*** in the event log ***E***. A model ***M*** is constructed by a set of nodes and arcs representing activities ***a*** and paths ***p,***

respectively. The path represents a trajectory from one node to its subsequent node.

Definition 2 (Event log). An event log ***E*** is a set of events (***c***, ***d***, ***t***). An event has a case identifier ***c***, a diagnostic code of ***d*** that happens at time ***t***. A trace ***T*** is an ordered sequence of events by timestamp ***t*** that happen to a case c, *where* $T \in E$.

Definition 3 (Pair log). A pair log ***P*** is a set of pairs of diagnostic codes where a pair event $\boldsymbol{\rho}$ has an identifier ***c***, an antecedent diagnostic code $\boldsymbol{d_i}$ and a subsequent diagnostic code $\boldsymbol{d_j}$ where $\boldsymbol{t_i}$ and $\boldsymbol{t_j}$ are the respective temporal information (i, j = 1, 2, 3, …, n, $\boldsymbol{\rho} \in \boldsymbol{P}$).

For each case, a trajectory of diagnostic codes is created. Three variables to construct the synthetic event log are *Case identifier*, *Activity*, and *Time stamps*. A case identifier is a unique value to represent a patient. An activity is a variable containing a diagnostic code following a formal coding standard as applied by health care organisation. A standard coding for diseases is used e.g., the ICD-9 or ICD-10. A time stamp contains information of when a certain diagnostic code occurred. Usually, this is the time of either the diagnosis has been made or being recorded. The time stamps facilitate the construction of a sequence of the diagnostic codes according to the trajectory directions. Table 3-1 presents a summary of the three variables to construct an event log.

Table 3-1 The sources of required data from the synthetic data set

Variables	**Data**	**Value example**
Case identifier	Patient identifier	“1234567”, “ABC123456”
Activity	Diagnostic code	“I21”, “J18”, “N17”, …
Time stamps	Time when the diagnosis was made or recorded	“26-02-2010”, “25-09-2014 15:45:17”

The term ‘selection of business process’ in this current stage’s activity is closely related to processes that commonly happen in a business organisation [10]. This thesis is analysing the pattern of disease occurrences using data in an electronic health record. For this current work, the term was translated into the selection of scope of analysis since there are more areas that can be analysed from an EHR. Therefore, for

the implementation of the method in case studies, the term ‘selection of business process’ was changed into ‘selection of the scope of analyses.

3.1.1.2 Defining research questions

The objective of this stage is to initiate the project by defining the research question(s). The input of this stage is a selected scope of analysis, and the outputs are research questions to analyse disease trajectories from a set of patients’ EHR.

A set of questions in the question-driven method was adopted in this current work. The method contains frequently-posed questions by medical professionals in process mining projects [50]. The adopted questions are:

1) What are the most followed paths and how long was the duration of the trajectories?

2) Are there differences in care paths followed by different patient groups?

3) Where are the long waiting time activities in the process?

The data set for this study has been checked for quality assessment by following a framework by Weiskopf & Weng [82] to check if they have a good quality to be used in process mining analysis.

3.1.1.3 Composing a team

A team was developed for the study by involving people from different backgrounds. The team consists of experts from the fields of computer science, cardiovascular medicine, epidemiology, and statistics. Collaborators from the health care domain have the role as the experts, according to the PM^2 methodology.

3.1.2 Stage-2: Extraction

The aim of this stage is to extract event data from the information system. The inputs are research questions and information systems, while the output is event data. This stage covers the activities of determining scope, extracting event data, and transferring process knowledge. The scope of the data extraction was defined based on the granularity of the data, the time, and the carefully selected attributes of the data. A synthetic data set was used to conduct a feasibility study in Section 3.2 and two data sets from the actual health information system were used in case studies which are discussed in Section 3.3 and more detailed coverage of both data sets are presented in

Chapter 4 and 5. The extraction part of the extract-transform-load is conducted at this stage while the transform and load are conducted in the next stage.

3.1.3 Stage-3: Data processing.

The stage's objectives are to create an event log and to prepare the event log for the mining and analysis stage. Following the data extraction in the previous stage, a construction of an event log and a series of transformation and filtering event log is required for mining disease trajectories. The result from those activities is a readily mined event log using process mining tools at the next stage.

The input for this stage is event data, and the output is the event log. There are four types of activities in this stage to produce the event log for the next stage:

- creating views
- aggregating events,
- enriching logs, and
- filtering logs.

A novel approach to identify disease trajectories is by combining a data-driven method by Jensen et al. (2014) [3] with the process mining method. Jensen et al.'s method is used to select the significant pairs of diagnostic codes and to determine the trajectory in each pair. The event log contains the events in a sequence, to allow the application of Jensen et al.'s method, event log transformations are needed, therefore 'transforming log' was added as part of the filtering logs activity. Detailed information on this approach is presented in the following sections.

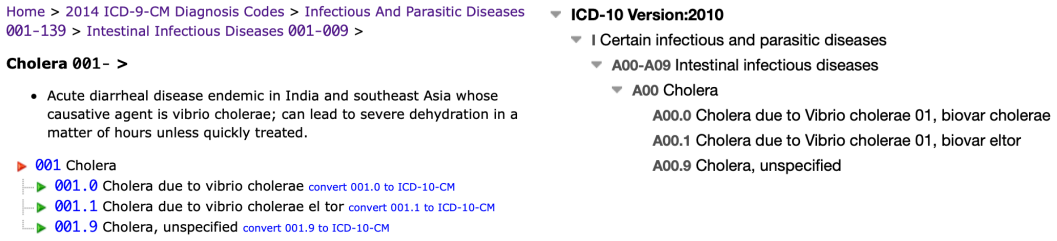
3.1.3.1 Creating views

The extracted event data and the aim of this research are the input of creating views activity. The patient was selected as the case notions, where the patient identifier becomes the case identifier. The occurrence of diagnostic code over time is selected as the event containing the activity name and timestamp. The temporal information was used to determine the sequence of the diagnostic codes. The event log is a compilation of events recorded in a format of *[case identifier, activity name, time]* and ordered by the case identifier and time.

3.1.3.2 Aggregating log

Diagnostic codes are selected to be the activity names of an event log. The aim of filtering activity is to reduce the complexity to focus on the analysis [10]. Health care organisations could adopt different standards for coding diseases.

One of the coding standards is the International Classification of Diseases (ICD) from the World Health Organisation (WHO). Two versions from the 9th revision (ICD-9) and the 10th revision (ICD-10) were used in this current work. Both the ICD-9 and ICD-10 use a hierarchical classification approach where diagnostic codes at the lower hierarchy represent more detailed information about the disease. Diseases coded with the lowest level of diagnostic codes can be aggregated to a level above them. This aggregation can be done until the highest level is achieved [106, 107]. Figure 3.2 below shows the hierarchy of both ICD-9 and ICD-10.



(a) (b)

Figure 3.2 Hierarchical structure of (a) the ICD-9 codes and (b) ICD-10 codes.

In this current work, the diagnostic codes were aggregated. The first case study in Chapter 4 and the second case in Chapter 5 used the ICD-9 and ICD-10, respectively. Both diagnostic codes in the case studies were aggregated into level-3, meaning only the first three characters from the complete code were taken into account. For example, a diagnostic code of *acute myocardial infarction of inferoposterior wall* in the ICD-9 code is 410.3 while in the ICD-10 the disease is coded as I21.1. Both codes were aggregated into level-3 became 410 and I21, respectively.

Aggregating diagnostic codes into level-3 may reduce the complexity of the analysis and focus on the primary aim to discover disease trajectories using process mining approach.

3.1.3.3 Filtering log

The aim of filtering activity was to produce an event log that contains patients' sequence of diagnostic codes where each sequence follows the below requirements:

1. contains frequently occurred diagnostic codes,
2. contains the first occurrence of each diagnostic code,
3. the diagnostic codes are strongly and significantly associated,
4. the associated diagnostic codes have a unidirectional trajectory.

To achieve the aim above, the filtering and transformation activities were done in sequence as follows:

1. Activity-based filtering
2. Event log transformation to pair log
3. Pair log filtering by frequency
4. Pair log filtering by Relative Risk
5. Pair log filtering by trajectory
6. Pair log transformation to event log

Details of each filtering and transformation activities are presented in the following descriptions:

1. Activity-based filtering

   Two activity-based filterings are needed to achieve an event log that only contains the first occurrence of frequently occurred diagnostic codes. First filtering is by applying the Pareto principle to select the frequent diagnostic codes from the selected event data and then followed by excluding the second or the repeating occurrence of diagnostic codes from each patient.

   a. Pareto principle [91]

   The Pareto Principle states that 80% of consequences come from 20% of the causes. This principle is a simplification of Pareto distribution [108]. The principle is a rule of thumb for general understanding, where most outcomes come from several causes. Applying the guideline within this study helps to focus on the impactful causes. For example, a study by Müller et al. (2014) [109] managed to test the Pareto principle using real-world data of emergency department admission due to adverse drug events (ADE). The result shows that the real data did not precisely follow the 80-20 but still supported the Pareto principle. Among 110 inpatient hospitalisations, 80% (n=88) of individuals ever encountered 35.8% (n=33) most frequent type of drugs that

caused ADE. The most relevant cause of ADE was identified using the Pareto Principle, allowing the clinician to develop a subsequent intervention to increase patient safety. Another benefit of applying the Pareto principle is to optimise the computing performance where any available resource can be maximised for exploration study.

b. Remove recurrent diagnostic codes

Any second or more recurrences of the same diagnostic code of each patient are removed and keep the first ever occurrence. This step is required to form the forward trajectories of a sequence of diagnostic codes. The removal is applied to the immediate or non-immediate reoccurrences.

For example, in Figure 3.3 (a), one patient has two immediate sequences of I25 that occurred on 02/03/2100 and 12/05/2100; in (b), a patient that has the same diagnostic code, but the second code occurred after another diagnostic code (I50). After the removal, the first occurrence of I25 for both patients was kept and the other was removed (see Figure 3.3 below).

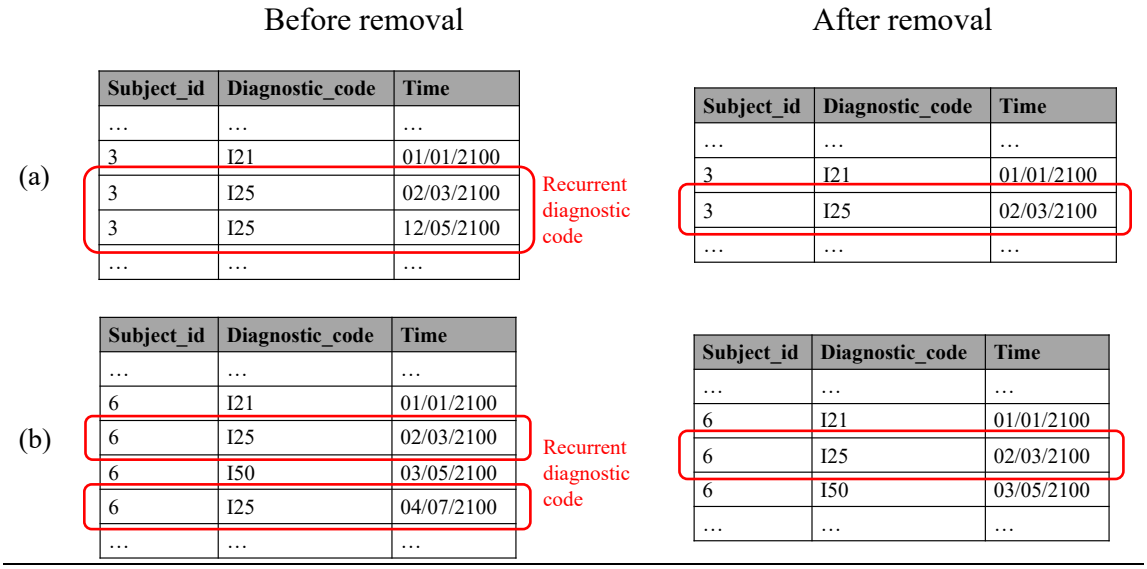


Figure 3.3 Filtering recurrent diagnostic codes of (a) immediate reoccurrence and (b) non-immediate reoccurrence.

2. Event log transformation into a pair log

The next step was transforming the filtered event log into a pair log. The transformation was required to allow filtering based on pairs. The immediate subsequent rows for each patient represent a sequence of diagnostic codes.

The pair is presented as two diagnostic codes (d_1 and d_2) and an arrow in between:

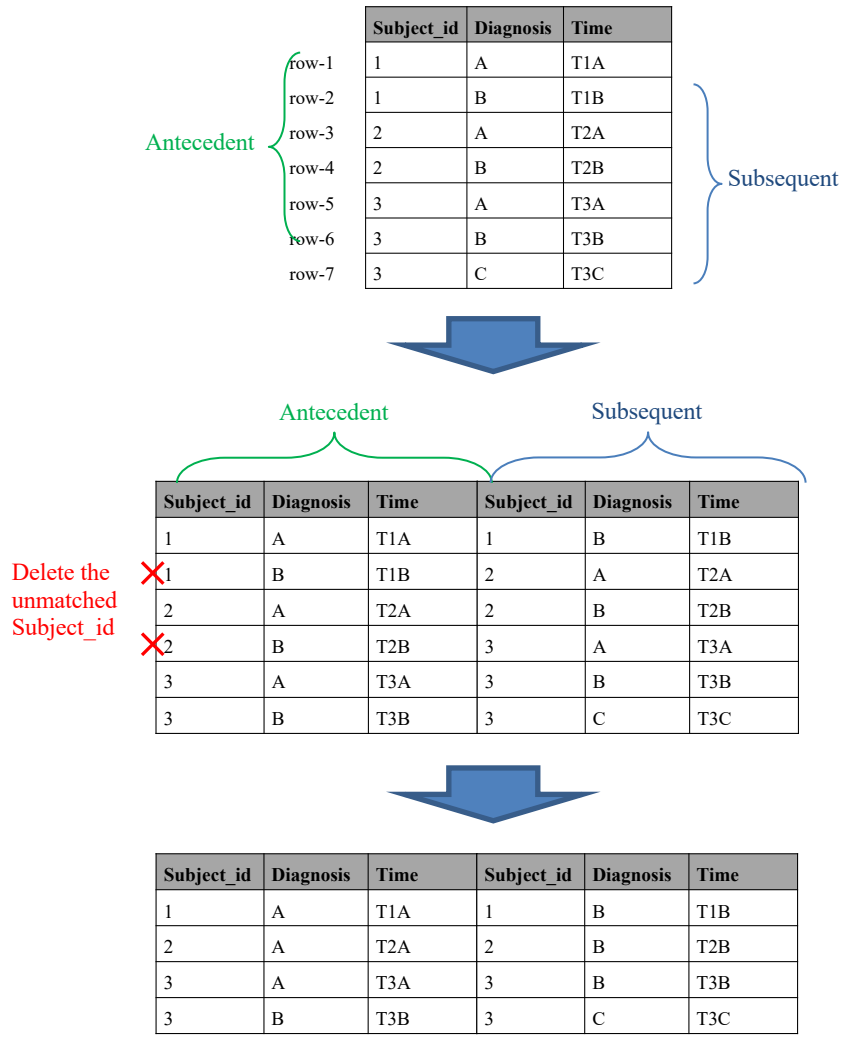
$$d_1 \rightarrow d_2$$

d_1 is the diagnostic code that occurred as the antecedent, and d_2 is the diagnostic code that occurred as the subsequent. A pair of I21 → I25 means d_1 = I21 and d_2 = I25.

To transform the sequence into pairs, the following algorithm can be followed:

Algorithm-1
Input: event log containing patient identifier, diagnostic code, and time. Output: pair log containing patient identifier, antecedent, time1, subsequent, time2.
Begin Select rows from the event log except the last row. Save the selection as the Antecedent data. Rename the columns of Antecedent data to indicate the selection will be placed at the left-hand side and the time become time1. Select row from the event log except the first row. Save the selection as the Subsequent data. Rename columns of Subsequent to indicate the selection will be placed at the right-hand side and the time become time2 Create pair log by concatenating the Antecedent and Subsequent. Drop any rows that have a different patient identifier of Antecedent and Subsequent. Drop the Subsequent patient identifier column of the pair log. End.

Figure 3.4 presents the example from a feasibility study in Section 3.2 showing the transformation an event log into a pair log containing three patients' data.



(c)

Figure 3.4 The process of transforming an event log of three patients; (top) row-1 until 6 are selected as the *antecedent* while row-2 until 7 will be the *subsequent*; (middle) the concatenation of antecedent and subsequent, rows with red crosses have unmatching Subject_id and removed; (bottom) the final construction of a pair log.

3. Pair log filtering by frequency

The aim of this filtering activity is to exclude pairs of diagnostic codes that only happened to one patient. The occurrence of each pair is calculated and then pairs with single occurrences are excluded.

4. Pair log filtering by Relative Risk [94]

Relative Risk (*RR*) is to measure the strength of the association of diseases in a pair. The measurement identified the likelihood of a subsequent disease (d_2) occuring after the antecedent disease (d_1). A pair with strong association is a pair that has an increasing likelihood of d_2 occurs after d_1 where the RR value is greater than 1 ($RR>1$). The *RR* comes with 95% confidence interval (CI). For any *RR* with

the CI including the value of 1.00 means that the RR is not significant since RR = 1.00 shows that there is no difference when d_2 is likely to occur after d_1 or vice versa. Pairs of diagnostic codes with $RR > 1$ and the lower CI is greater than 1 are taken for the next filtering activity.

An RR value is calculated using a contingency 2 × 2 table presented in Table 3-2.

Table 3-2 Contingency 2 × 2 table.

	Case	Controls
Exposed	a	b
Unexposed	c	d

Using the variables in the table above, the relative risk is computed using the following formula:

$$RR = \frac{\frac{a}{(a+b)}}{\frac{c}{(c+d)}}$$

where:

a is the number of patients with d_1 and followed with d_2;

b is the number of patient with d_1 but **not** followed with d_2;

c is the number of patients without d_1 but followed d_2; and

d is the number of patients **without** both d_1 and d_2.

The RR measurement above is to get the strength of the incidence of subsequent diagnostic code to the antecedent. The distribution of RR is unknown regardless of the sample size, but the distribution of the natural log of RR is approximately a normal distribution, therefore the measurement of 95% confident interval for relative risk should be done in two steps [110]. First, a confidence interval for the RR is generated by computing the natural log of RR:

$$\ln(RR)$$

as it is approximately having a normal distribution. Second, by computing the antilog of both the upper and lower limits of the confidence interval for the above natural log of RR. The computation follows the below formula:

$$ln(RR) \pm z\sqrt{\frac{(b/a)}{(a+b)} + \frac{(d/c)}{(c+d)}}$$

$$\Leftrightarrow ln(RR) \pm 1.96\sqrt{\frac{(b/a)}{(a+b)} + \frac{(d/c)}{(c+d)}}$$

where z is equal to 1.96 the limit number where 95% of standard deviation considered as significant.

The resulting confidence interval values are for the natural log of *RR*, therefore the antilog of both the upper and lower limit resulting the 95% confidence interval for the *RR*:

$$\left(Exp\left\{ ln(RR) - z\sqrt{\frac{(b/a)}{(a+b)} + \frac{(d/c)}{(c+d)}} \right\}, Exp\left\{ ln(RR) + z\sqrt{\frac{(b/a)}{(a+b)} + \frac{(d/c)}{(c+d)}} \right\} \right)$$

The next step is to measure if the association of the paired diagnostic codes $d_1 \rightarrow d_2$ is statistically significant by testing the independence of the two categorical variables: the incidence of d_1 and the incidence of d_2.

In this current work, the Karl Pearson chi-square test was used to determine if the paired diagnostic codes of d_1 and d_2 are associated or just occurred by chance. The statistical question was "is the incidence of d_2 related to d_1?" In another way, two hypotheses were presented to be tested:

Null hypothesis/ Hypothesis$_0$ (H_0):

The incidence of d_1 and d_2 are independent.

Alternate hypothesis/ Hypothesis$_A$ (H_A):

The incidence of d_1 and d_2 are associated.

To answer the question or to test the hypothesis above, a contingency table is used to put the observed frequency of d_1 and d_2 – where both values are taken from the data set, and to calculate the expected frequency when the incidence of both d_1 and d_2 are by chance – i.e., we expect that there is no association between d_1 and d_2. The aim of this test is to see if the observed frequency is no different to the expected frequency. The Karl Pearson Chi-Square test will be addressed as "Chi-Square" from this point forward.

Using the same contingency table as shown in Table 3-2, the contingency table for the Chi-Square test is used to calculate the expected frequency. The calculation of the expected frequency (E) for each cell in Table 3-2 follows the below concept:

$$Expected\ frequency = \frac{Row\ Total\ \times Column\ Total}{Grand\ Total}$$

An example formula of calculating expected frequency for cell-1 containing an observed value (O) of *a* is:

$$E_1 = \frac{(a+b)\times(a+c)}{(a+b+c+d)}$$

Table 3-3 shows the expected frequency calculation for all cells in the contingency table.

Table 3-3 Expected frequency calculation in a contingency table.

	Case	Control
Exposed	$E_1 = \frac{(a+b)\times(a+c)}{(a+b+c+d)}$	$E_2 = \frac{(a+b)\times(b+d)}{(a+b+c+d)}$
Unexposed	$E_3 = \frac{(c+d)\times(a+c)}{(a+b+c+d)}$	$E_4 = \frac{(c+d)\times(b+d)}{(a+b+c+d)}$

The observed and the expected frequencies are then used to obtain the Chi-Square value using the following formula:

$$\chi^2 = \sum_{i=1}^{n} \frac{(\mathrm{O}_i - \mathrm{E}_i)^2}{\mathrm{E}_i}$$

where:

n is the number of cells in the table,

O is the observed frequency,

E is the expected frequency.

A single Chi-Square values from the above formula shows the difference between the observed frequencies and the expected frequencies. The bigger the Chi-Square value reflects the bigger difference of observed and expected frequency.

Next step is to calculate the *degree of freedom* (DF) as required to find out if the Chi-Square value is statistically significant to reject the null hypothesis. The DF is calculated as follows:

$$DF = (r - 1)(c - 1)$$

were

r is the number of rows of the contingency table, and

c is the number of columns of the contingency table.

Referring the contingency table in Table 3-2, it has two rows (r=2) and two columns (c=2). These values are then used to obtain the DF:

$$DF = (2 - 1)(2 - 1) = 1$$

By having the Chi-Square value and the DF, the next step is determining the probability level. A common convention of probability level of 0.05 is taken (p = 0.05). The value reflects the 5% chance or less where the observed difference is due to chance. Using the Chi-Square value and the DF then the critical limit of 5% can be obtained by consulting the table of "critical values of Chi-Square" [96]. The critical values can be found by the intersection of DF and the probability value which is 3.84 (Figure 3.5 shows how to find a critical value using the table of "critical values of Chi-Square"). If the Chi-Square value is greater than 3.84 (p-value < 0.05) then it suggests that the occurrence of d_2 was unlikely by chance and the null hypothesis (H_0) can be rejected. In other words, there is a temporal association between the paired diagnostic codes $d_1 \rightarrow d_2$.

d.f.	.995	.99	.975	.95	.9	.1	.05	.025	.01
1	0.00	0.00	0.00	0.00	0.02	2.71	3.84	5.02	6.63
2	0.01	0.02	0.05	0.10	0.21	4.61	5.99	7.38	9.21
3	0.07	0.11	0.22	0.35	0.58	6.25	7.81	9.35	11.34
4	0.21	0.30	0.48	0.71	1.06	7.78	9.49	11.14	13.28
5	0.41	0.55	0.83	1.15	1.61	9.24	11.07	12.83	15.09
6	0.68	0.87	1.24	1.64	2.20	10.64	12.59	14.45	16.81
7	0.99	1.24	1.69	2.17	2.83	12.02	14.07	16.01	18.48
8	1.34	1.65	2.18	2.73	3.49	13.36	15.51	17.53	20.09
9	1.73	2.09	2.70	3.33	4.17	14.68	16.92	19.02	21.67
10	2.16	2.56	3.25	3.94	4.87	15.99	18.31	20.48	23.21
11	2.60	3.05	3.82	4.57	5.58	17.28	19.68	21.92	24.72
12	3.07	3.57	4.40	5.23	6.30				26.22
13	3.57	4.11	5.01	5.89					
14	4.07	4.66							

Figure 3.5 An excerpt of critical values of Chi-Square (adapted from [96]) to find a critical value of DF=1 and probability value 0.05.

The implementation of the Chi-Square test in this current work was done automatically in Python by incorporating a statistical package named "SciPy" [111].

We will now describe the criteria for selecting the paired diagnostic codes being advanced for the directionality test:

- RR value more than 1 ($RR > 1$), means that the risk of having d_2 after having d_1 is strongly likely; and
- the lower limit of the confidence interval is more than 1 since RR = 1 means that the risk of d_1 and d_2 are equal, or can't be determined: and
- the p-value of the Chi-Square test is less than 0.05 ($p < 0.05$), meaning the paired diagnostic codes are significantly associated.

The selection of paired diagnostic codes using the above criteria may still include paired diagnostic codes of $d_1 \rightarrow d_2$ and vice versa $d_2 \rightarrow d_1$. The next section determines a single direction of paired diagnostic codes to build the trajectory and to avoid a loop in the disease trajectory model.

5. Pair log filtering by trajectory

Using a collection of statistically significant paired diagnostic codes in the pair log, then a directionality test is conducted. A binomial test is used to determine if a disease d_1 occurred before d_2, or vice versa [112]. The aim of filtering the pair log by directionality is to have a pair log that has unidirectional diagnostic pair sequences – i.e., only one diagnostic pair sequence of $d_1 \rightarrow d_2$ and $d_2 \rightarrow d_1$ will be considered.

A hypothesis testing is needed to determine if the diagnostic pair sequence is statistically significant. There are three conditions of the diagnostic pair sequences that may occur:

(A) Only one out of two diagnostic pair sequences of $d_1 \rightarrow d_2$ and $d_2 \rightarrow d_1$ is significant as a trajectory.

(B) Both diagnostic pair sequences of $d_1 \rightarrow d_2$ and $d_2 \rightarrow d_1$ are non-determinable (either both pairs are significant or non-significant).

(C) The diagnostic pair sequence of $d_1 \rightarrow d_2$ is not significant as a trajectory.

The disease trajectory model is expected to only contain diagnostic pair sequences with one directionality, therefore a hypothesis testing is needed to determine which of the three conditions above is achieved. The hypothesis testing used in this study is the Binomial test using a significant level of 0.05 and a probability of 50%. A Python package named SciPy is used for the binomial test.

Due to the nature of the exploratory study undertaken for this current work, no adjustments were made for multiple testing of the Relative Risk measurement and the binomial test [113]. This was taken to avoid accepting the null hypothesis (H_0) when H_0 is false (type II errors, also known as 'false-negative') [114].

The final pair log contains a collection of diagnostic pair sequences that have a statistically significant association measurement (RR > 1) and dominant directionality. A retransformation from a pair log into an event log is required to allow process mining analysis to be undertaken. This transformation is explained in the next step.

6. Pair log transformation to event log

This step is to transform the pair log into an event log to make the data set ready for process mining analyses. The retransformation of a pair log into an event log can be done by following the below algorithm:

Algorithm-2
Input: pair log containing patient identifier, antecedent diagnostic codes, time1, subsequent diagnostic codes, and time2. Output: event log containing patient identifier, diagnostic code, and time.
Begin Select columns of the pair log containing patient identifier, antecedent diagnostic codes, and time1 into a new table. Rename the column of antecedent diagnostic code and time1 to other names that isn't showing the diagnostic code's position when in the form of pairs. Select columns of the pair log containing patient identifier, subsequent diagnostic codes, and time2 into a new table. Rename the column of subsequent diagnostic code and time2 to other names that isn't showing the diagnostic code's position when in the form of pairs.

Combine the above new tables by stacking the tables to create a single table. Order the data by patient identifier and time. End.

The ordering of the event log is important since process discovery algorithms sequentially read the rows of event data in an event log, starting from the first row until the last. This method of reading events reproduces non-significant pairs of diagnostic codes that have been excluded in the filtering log step. Excluded diagnostic pair sequences create new 'connections' between the remaining diagnostic pair sequences and the new connections have the possibility of recreating the pairs that are already measured as non-significant. For example, one patient had a sequence of six diagnostic codes: I25 → I24 → AMI → I50 → J22 → J18. The diagnostic pair sequence in the pair log become: I25 → I24, I24 → AMI, AMI → I50, I50 → J22, and J22 → J18. After the filtering log, two diagnostic pair sequences are measured as less significant: I24 → AMI and I50 → J22. the exclusion of less-significant pairs creates a pair log of I25 → I24, AMI → I50, and J22 → J18. Retransformation of a pair log into an event log creates the sequence of: I25 → I24 → AMI → I50 → J22 → J18. The final sequence shows that the filtering log still produces a similar sequence of diagnostic codes as before the filtering log activity. Figure 3.6 illustrates the reoccurrence of the non-significant pairs of diagnostic codes after the transformation of a pair log into an event log.

To handle the above situation, an additional checking to the recreation of the less-significant diagnostic pair sequences is required. Number of patients affected by this recreated pairs needs to be calculated and then consult the result with the expert to address the issue. To calculate the number of patients who are affected by the reoccurrence of less-significant diagnostic pair sequences can be done by following the below algorithm:

Algorithm-3
Input: Event log as a result of retransformed pair log after filtering log activities.

Output: Number of patients who affected with the reoccurrence of less-significant diagnostic pair sequences, including the percentage from the total number of patients in the final event log.
Assumptions: - Relative risk, Chi-Square test, and Binomial test were performed. - The first transformation of event log to pair log and the retransformation from pair log into event log were performed. - Excluded diagnostic pair sequences are recorded and identifier to each pair were assigned.
Begin Transform the event log into pair log. Add new column as a marker to the diagnostic pair sequences that also exist in a table containing the excluded diagnostic pair sequences. Change the diagnostic codes of the affected pairs into any names to show the diagnostic codes as part of the non-significant pairs. For example, 'any non-significant traces' (ANST). Retransform the pair log into event log where the altered diagnostic codes are taken. Count the number of patients who have 'ANST' in the sequence. Count the percentage of patient who have 'ANST' in the sequence. End.

If the number of the affected patient is considered as small enough compared to the overall number of patients in the final event log, then exclusion is considerable. A case study in Chapter 5 shows how this issue has been addressed.

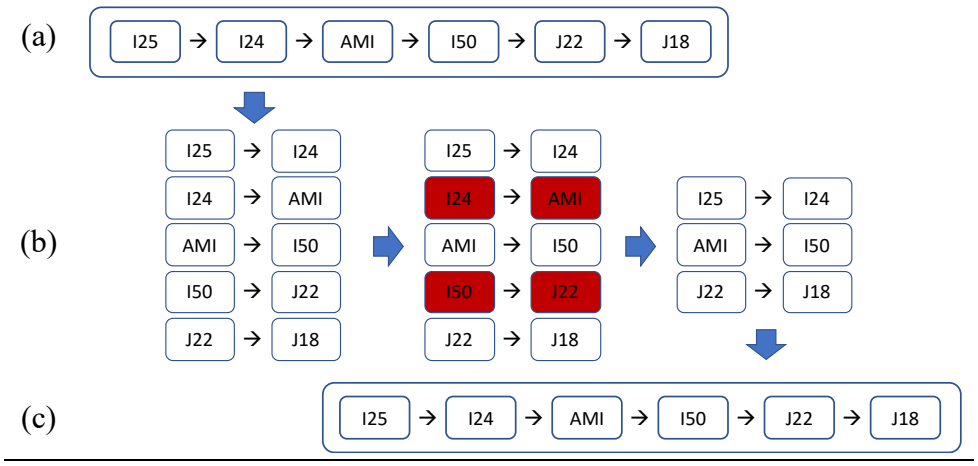


Figure 3.6 The reoccurrence of excluded diagnostic pair sequences after a retransformation of a pair log into an event log.

The steps undertaken at this stage are based on discussions with the experts to get a collective agreement on the scenario of filtering activities, to get their expert judgment regarding their specialities in medicine or statistical analyses, and to share their knowledge. Multiple iterations in filtering log activities may be undertaken until an agreement is achieved.

Filtering the log using the Relative Risk measurement, Chi-Square test and the Binomial test is useful to separate the significant and non-significant diagnostic pair sequences. The non-significant diagnostic pair sequences are worth for further analyses to get a complete view of mining disease trajectories using EHR.

3.1.3.4 Enriching log

Alongside the three minimum requirements to construct an event log – patient identifier, diagnostic codes, and time stamps, more information related to the patient can be added as attributes. The aim of enriching the event log is to allow post-hoc analyses based on the available attributes. For this current work, three attributes of sex, age group, and mortality status were added into the event log.

3.1.4 Stage-4: Mining and analysis.

In this stage, process mining techniques were applied with the objective of answering the research questions and gathering understanding. The input and output of this stage are event logs and findings to answer the research questions. Four activities in this stage are:

1. process discovery,
2. conformance checking,
3. enhancement, and
4. process analytics.

3.1.4.1 Discovery

Process discovery in process mining is to generate a process model using a certain algorithm where event log is the input. The discovery for this disease trajectory study also used algorithms that are intended to discover process model, but since the study is about disease trajectory model, then the discovered model is called the disease trajectory model from this point onwards.

Decision on the discovery algorithms in this current work was based on the characteristics of the data and the purpose of the analysis. Disease trajectory mining involves using data containing sequences of various lengths of events and noise may still be exist in the event log in the form of infrequent data. The purpose of this study is to identify long trajectories to reveal the behaviour and robust to the noise. Based on the above, the main algorithm in this study is the Heuristics Miner. The algorithm is tolerant to noise and has the technique to search long distance dependency relations [115]. The algorithm is available as a plugin named "interactive Data-aware Heuristics Miner" (iDHM) in one of the process mining tools named ProM Framework version 6.8 [52, 53].

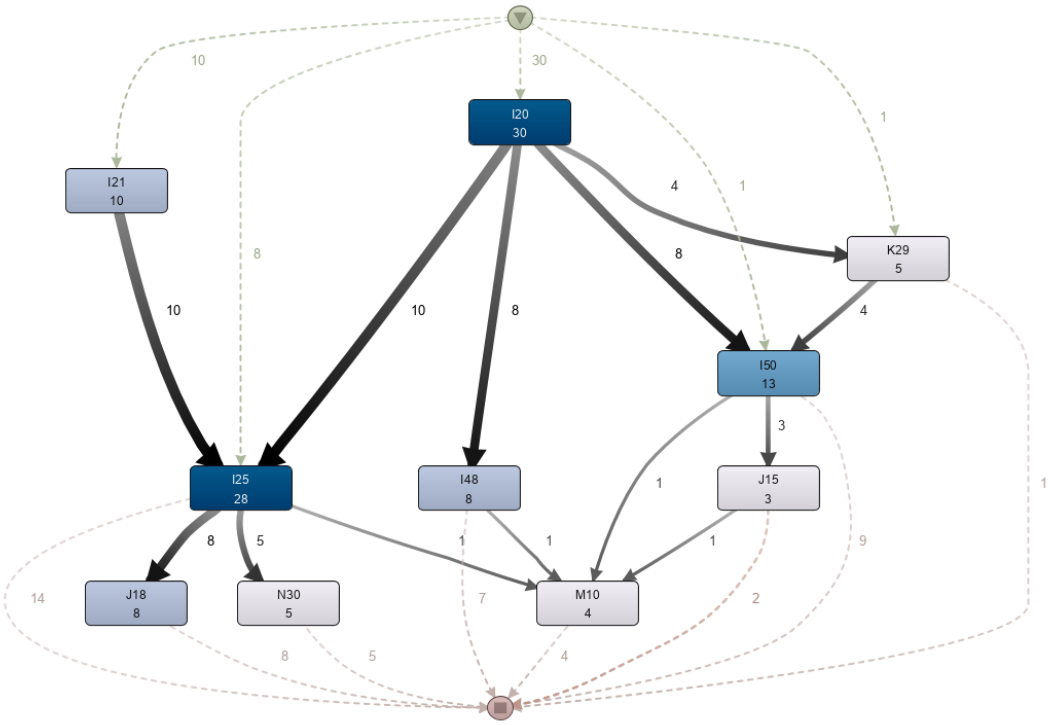
Other process mining tools for discovering disease trajectory models are DISCO [55, 116], and Celonis [56]. ProM 6.8 provides various discovery algorithms such as the Heuristics Miner and the Inductive Miner while DISCO and Celonis use the extended version of the Fuzzy Miner as the one only discovery algorithm available.

The result of discovering disease trajectories was mainly presented using the disease trajectory model, trace variants, and other supplemental figures to support the specific analysis. The main visualisations in this study are described as the following.

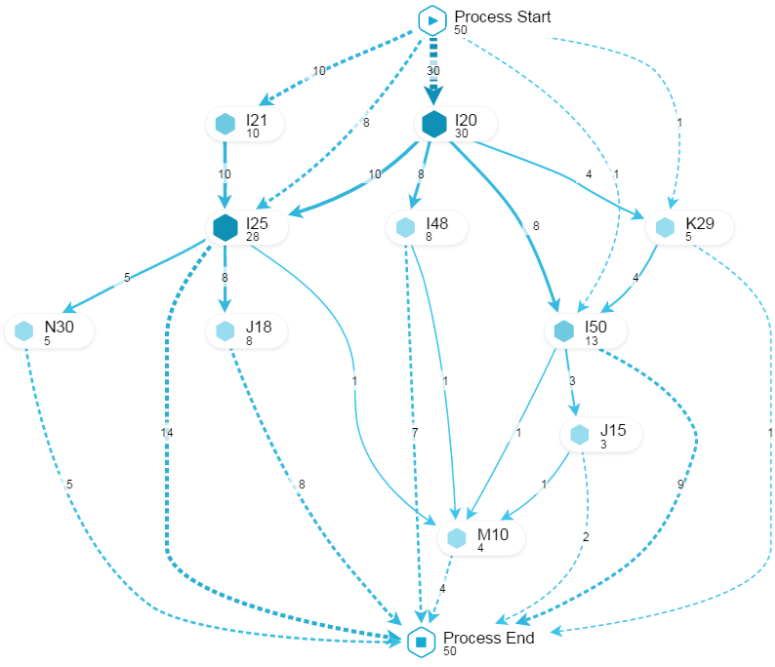
1. Disease trajectory model

A disease trajectory model represents a sequence of events as described in Definition 2 in section 3.1.1.1. The type of visualisation for a disease trajectory model is a directly-follows graph composed by nodes and directed arrows. The concept of this type of visualisation is used by the iDHM plugin in ProM 6.8, DISCO, and Celonis despite the minor differences in presentations.

Examples of disease trajectory models for each process mining tools are presented in Figure 3.7, where a disease trajectory model is taken from a feasibility study in section 3.2 generated using DISCO (a) and Celonis (b).



(a)



(b)

Figure 3.7 Examples of disease trajectory models; (a) a model generated from DISCO, while (b) was from Celonis.

The two models above are generated from the same event log. The layout and the shape of the nodes may look different, but both are showing the same behaviour of the event log.

2. Trace variants

A trace variant is used to represent the sequence of events including how many cases are following the patterns. Figure 3.8 shows the example of the five most frequent trace variants adapted from Chapter 5, generated using ProM 6.8. The trace variants visualisation contains the identified patterns where each activity in the trace is colour coded to increase readability. It also displays the number of cases that followed the trace pattern including the proportions from the overall data.

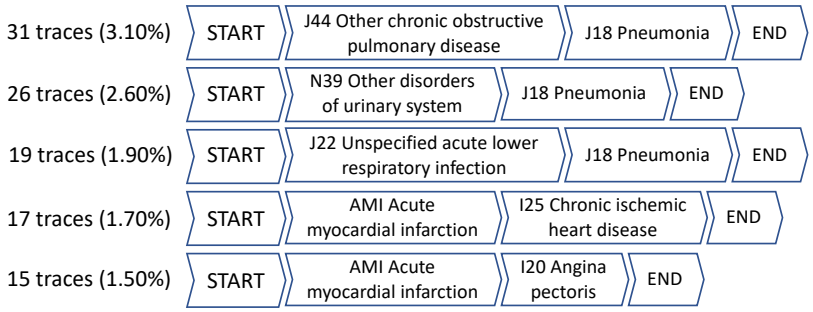


Figure 3.8 Trace variants example.

3.1.4.2 Conformance checking

Conformance checking is done using ProM 6.8 since it provides a richer set of measures in the form of plugins, compared to Celonis. Below is the list of plugins in ProM 6.8 that were used in this study:

1. *Replay a Log on Petri Net for Conformance Analysis*

This plugin helps to conduct conformance checking and provides fitness measures [117, 118]. The measurement is done by replaying the data from event log into the generated model. The number of skipped activities and the inserted activities were computed to determine the fitness value. Skipped activities are activities available in the model but in some cases those activities cannot be played as they are missing in the event log. Inserted activities are activities available in the log, but the model prevents them from happening.

2. *Check Precision based on Align-ETConformance*

This plugin helps in measuring the precision of the discovered disease trajectory model. Precision is to determine how accurate the model is at representing the

behaviour from the log but not underfitting. The measurement of precision using this plugin is chosen since the ability to accurately measure precision of non-completely fitting traces in the log, hence provide a more robust approach [43, 119]. This plugin is the improvement of the previous precision measurement [40] from the same authors.

3. *Measure Precision/ Generalization*

This plugin is to measure a generalisation of a model where it still be able to accept various patterns of future traces. Although this plugin has the property to measure precision it also has a measurement of generalisation [40]. Only the value of the generalisation was considered in this study since the precision measurement approach has been replaced with the updated version.

The above plugins for conformance checking were used in the feasibility study (section 3.2) and the case studies in Chapter 4 and Chapter 5. Otherwise stated, only some of the plugins were used depends on the goal of the analysis.

3.1.5 Stage-5: Evaluation.

The evaluation stage has the objective to translate the findings in the previous stages into a meaningful contribution whilst achieving the goals of the study. Two activities in the PM^2 framework were followed and described in the following sections.

3.1.5.1 Diagnose

Obtained findings through mining & analysis need to be diagnosed through the following activities:

1. Understanding the disease trajectory model to allow correct interpretation to the results.
2. Highlighting interesting result or unexpected result if any.
3. Detecting for further research questions or refining the current research questions for possible iterations.

The iterations may affect Stage-2, Stage-3, and Stage-4, or any combinations of the three stages. The above activities are done collaboratively with the experts with aim to share the knowledge among the team.

3.1.5.2 Validation and verification

The obtained process model from the previous steps needs to be evaluated for representativeness to the event log. To evaluate the representativeness of the obtained process models, validation and verification steps are needed.

1. Validation

   In this step, the conformance checking measurement is used as an indicator to evaluate the discovered process model using a validation data. The validation data is a subset of the complete event log but is never used to discover the process model. The rest of the event log is used as training data to discover process model and then conformance checking is performed using the discovered model against the validation data. This process is called cross-validation.

   The conformance checking results from Stage-4 are evaluated using *k-folds cross-validation*. The readily loaded event log into process mining tool is divided equally into *k* parts of sub-event log by random, and then use one part of sub-event log as the validation event log while the rest of the part were used as training data.

   A disease trajectory model was generated using the training data and used it for conformance checking against the validation data. Parameters used for mining disease trajectories should remain similar for each training data to make the cross-validation as a fair process. The cross-validation was done *k* times to allow each sub-event log to have a chance once to be the validation data.

   The average value of each replay fitness value, generalisation value, and precision value is then compared against the conformance checking results using the full-sized event log. The difference of the average values from the cross-validation should be close enough with the values from using the full-sized event log. Average values of conformance checking from the cross-validation are anticipated to be lower than the conformance checking using the full-sized event log. There are possibilities where some diagnostic codes in the validation data are not available in the training data or vice versa.

2. Verification

Obtained findings from the data analysis are communicated through discussions with the experts to receive feedback and to share the knowledge. In case any unexpected findings obtained, they can be consulted to experts for correctness.

The validation process is useful to test the stability of the discovered disease trajectory models. If the conformance checking results using the validation data are high, then it is fair to say that the discovered model is in good quality and can represent the actual behaviour.

The implementation of the validation and verification method was conducted in case studies that presented in Chapter 4 and 5. A conference paper showing the method implementation has been presented and published [86]. Validation and verification are important to conclude if the mined disease trajectories are robust and representative to the data.

3.2 Feasibility study

Before implementing the method of mining disease trajectories using actual EHRs, a feasibility study of the method was conducted using a synthetic data set. A feasibility study was conducted as an assessment to process mining for identifying and constructing disease trajectory model. PM^2 framework was followed in this current work and a subset of a disease trajectory model by Jensen et al. (2014) [3] was used as a reference to recreate the synthetic event log.

3.2.1 Stage-1: Planning

In the early stage of this study, a literature review of process mining in cardiology was conducted and found that there was no article reporting the use of process mining to understand disease trajectories [70]. A different approach to identifying disease trajectories has been made by Jensen et al. [3] who looked at the temporal aspect of the diseases' occurrences and applied some statistical techniques to build the trajectories. Jensen et al. defined a disease trajectory as a temporal order of diagnostic codes that observed in the patients.

Jensen's disease trajectory model was generated using paired diagnostic codes of 6.2 million individuals extracted from the Danish National Patient Registry, which includes data on all hospital encounters between 1996 and 2010. Nearly 1.2 million

pairs of diagnostic codes were identified, where each pair represents a time-dependent sequence association. The 1.2 million pairs were then condensed to 4,014 pairs with a significant correlation and directionality. The final collection of diagnostic code pairs was combined to form longer trajectories of three or more diagnostic codes. Figure 3.9 shows the final disease trajectory model by Jensen et al. (2014). In the figure, the node represents the diagnostic codes, while the arrow represents the direction of the trajectory. The arrow's thickness represents the number of patients who followed the trajectory.

This current work used synthetic data where the event log was created using the subset of Jensen's disease trajectory model as a reference. The event name was taken from the node, and the temporal information was taken from the connected pairs. The antecedent and the subsequent represent which disease comes first before another. Finally, the thickness of the arrow represents the frequency of the trajectory. The synthetic data was used as input for the process mining tool, and the produced model was examined for similarity to the reference model.

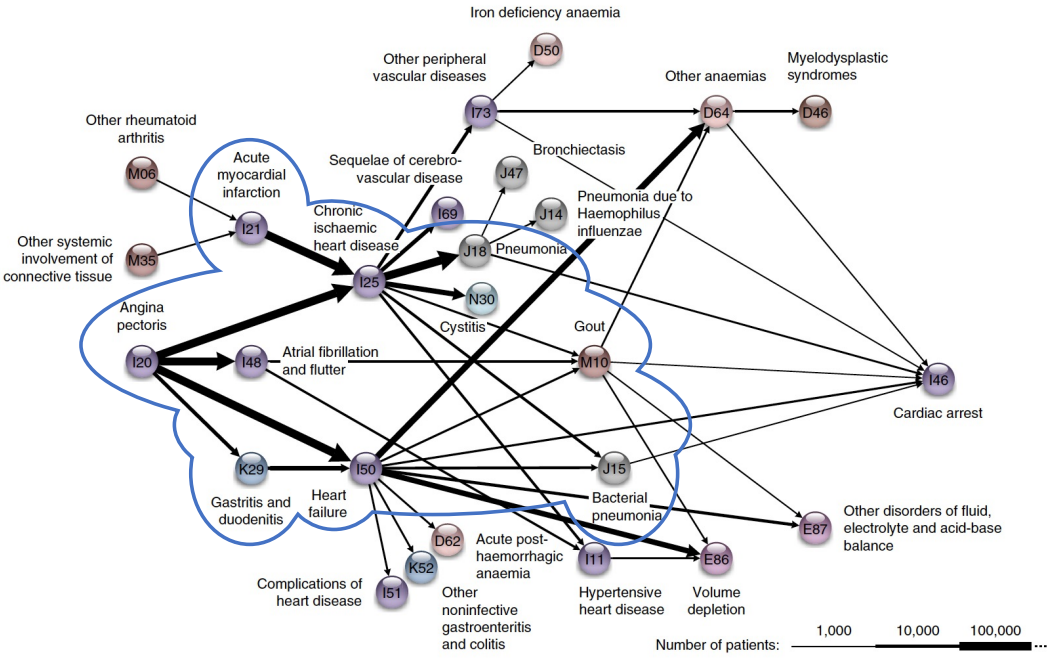


Figure 3.9 Disease trajectory model taken from Jensen et al. (2014) [3]; the selected trajectories are within the blue line.

The properties of the disease trajectories are similar to the property of business processes as explained in the introduction Section 1.2. Therefore, the size and coverage of the disease trajectories depend on the size and the coverage of the data

source – the electronic health records that at least contain patient identification, diagnostic codes, and time. In this feasibility study, an event log was created from a subset of a disease trajectory model produced in [3] (see Figure 3.10). The diagnostic codes were treated as the corresponding of activity in process mining and the directed arc corresponding of the sequence. A synthetic event log was planned to be constructed with minimum components of patient identification, diagnostic codes, and timestamp.

A research question was identified as *by having an event log where diagnostic codes correspond as the activity and the sequences defined by timestamp, can disease trajectories be identified using a process mining approach?*

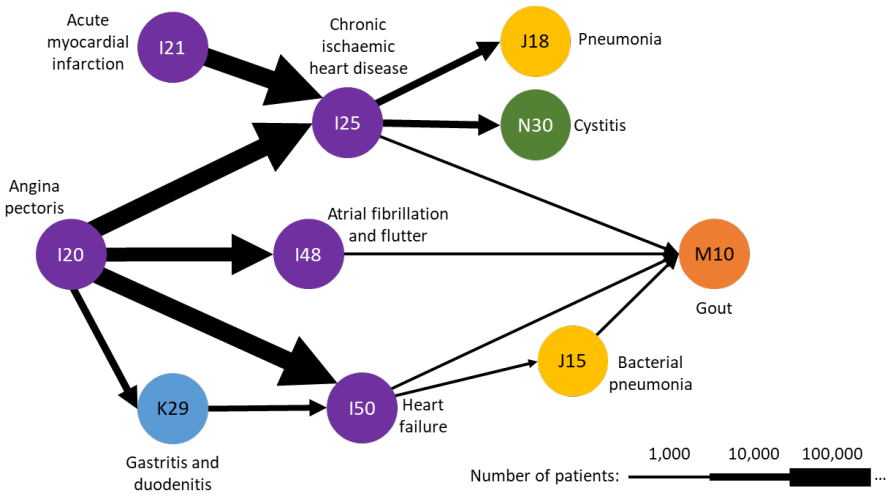


Figure 3.10. A subset of disease trajectory model extracted from a disease trajectory adapted from Figure 4.b in [3].

3.2.2 Stage-2: Extraction

Two extractions were conducted for this current work. First, a subset of disease trajectory was extracted as a reference to create an event log; second, a simulated extraction of event data from electronic health records by creating synthetic event data using the subset of disease trajectory as a reference. From the disease trajectory model by Jensen et al. (2014) (Figure 3.9), some trajectories were chosen arbitrarily involving various diagnostic codes, proportions, and lengths. Following the first extraction, an event log was created using the selected properties and the characteristics were maintained in the data.

1) Extraction of a disease trajectory model

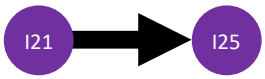
A new disease trajectory model for this feasibility study was drawn from the selected nodes and arcs within the area under the blue line in Figure 3.9. The result of the selection is presented in Figure 3.10.

A trajectory was identified as a bigram where each trajectory contains pair of consecutive diagnostic codes. The subset of trajectory in Figure 3.10 contains various trajectories at various lengths. A trajectory of I20→I25 means the diagnostic code of I20 temporally occurred before the diagnostic code of I25. In the figure, the trajectory of I20→I25 also has a subsequent diagnostic code of J18 to form a trajectory with the length of three: I20→I25→J18. Jensen et al. constructed the trajectory from pairs by overlapping similar diagnostic codes like the domino card game without checking if there are any cases that have such patterns. In contrast with the process mining way, the sequence of the diagnostic codes was taken from the electronic health record that represents the factual condition. Since the trajectory was identified as pairs, then the mentioned example contains two trajectory pairs of I20→I25 and I25→J18.

The thickness of the arcs reflects the number of patients who followed the trajectory. To simulate the thickness, any case containing the trajectory was added multiple times manually, until the proportion of the ratio was fulfilled. Trajectories with various lengths were also added to reflect the actual pattern of disease progression. Figure 3.11 illustrates the creation of the event log.

2) Creating a synthetic event log

Figure 3.11(a) shows how the event data were created following the selected trajectory. Timestamps were created to represent a trajectory sequence, and multiple cases for the same trajectory were added to represent the thickness of the arcs. The trajectory of I21→I25→J18 was translated into three rows of data as presented in Figure 3.11(b).



Subject_id	Diagnostic_code	Time
3	I21	01/01/2100
3	I25	02/03/2100
4	I21	21/02/2100
4	I25	14/06/2100
6	I21	02/01/2100
6	I25	03/01/2100
…	…	…

(a)

Subject_id	Diagnostic_code	Time
1	I21	01/01/2100
1	I25	02/01/2100
2	I21	02/01/2100
2	I25	02/06/2100
3	I21	01/01/2100
3	I25	02/03/2100
…	…	…
6	I21	01/01/2100
6	I25	02/01/2100
6	J18	03/01/2100
…	…	…

(b)

Figure 3.11 Creating event log from a subset of a disease trajectory by Jensen et al. (2014). (a) Events of a trajectory I21àI25 from multiple cases were added into event log; (b) Events of a trajectory I21→I25→J18 was added.

3) Adding noise to the event log

Disease trajectory is defined as the sequence of diseases that occurred for the first time over a period of time. Patients may have a recurrent disease in reality and recorded as multiple diagnostic codes (e.g., I21→I25→I25). The disease may also reoccur after other diseases (e.g., I21→I25→J18→I25→I25. This type of pattern was added as noise to the event log to simulate the reality.

Recurrent disease was translated into multiple rows of the same diagnostic codes including timestamps to represent the sequence. This noise addition is illustrated in Figure 3.12.

Subject_id	Diagnostic_code	Time
…	…	…
3	I21	01/01/2100
3	I25	02/03/2100
4	I21	21/02/2100
4	I25	14/06/2100
…	…	…

(a)

Subject_id	Diagnostic_code	Time
…	…	…
3	I21	01/01/2100
3	I25	02/03/2100
3	I25	12/05/2100
4	I21	21/02/2100
4	I25	14/06/2100
…	…	…

Recurrent diagnostic code

(b)

Figure 3.12 (a) Event log before noise addition and (b) after the noise addition.

4) Final synthetic event log

The final event log contains 126 rows of event data out of 50 cases, including the noise that had been added. This final event log is available online in GitHub [120] to allow other process miners to reproduce the method and the analysis for the feasibility study. The event log only contains the minimum requirement of case id, activity, and timestamp. The development tool of synthetic event log was Microsoft Excel, where the data were curated manually to reflect the subset event log as presented in Figure 3.10.

3.2.3 Stage-3: Data processing

Data processing for the feasibility includes creating views, filtering event log, and transforming event log into pair log. The relative risk and the binomial test were not conducted in this feasibility study as the aim is to apply process mining method using event log containing event data of diagnostic codes, performing event log transformations to reproduce a published disease trajectory model. Also, conducting simulation of relative risk calculation and the binomial test requires information on the actual proportion of each diagnostic code's occurrence relative to its antecedence. This information is not completely provided in Jensen se al.'s published paper nor in its supplementary work.

The view created in this stage was to produce an event log that followed the format of *[subject_id, activity, time]*. The *activity* field contained the diagnostic codes and therefore the field name was change into *diagnosis* to made it more intuitive. The event filtering was done by removing recurrent diagnostic codes, only the first occurrence of each diagnostic for each patient were retained. Following the event log filtering, a transformation of event log into a pair log was done. Figure 3.13 shows the

illustration of how the recurrent diagnostic code was removed and how the event log being transformed into a pair I.

Subject_id	Diagnosis	Time
1	I21	01/01/2100
1	I25	31/01/2100
2	I21	02/01/2100
2	I25	02/06/2100
3	I21	01/01/2100
3	I25	02/01/2100
3	I25	03/01/2100
…	…	…

Recurrent diagnostic code

(a)

Subject_id	Diagnosis	Time
1	I21	01/01/2100
1	I25	31/01/2100
2	I21	02/01/2100
2	I25	02/06/2100
3	I21	01/01/2100
3	I25	02/01/2100
…	…	…

(b)

Subject_id	Antecedent	Subsequent	Time1	Time2
1	I21	I25	01/01/2100	31/01/2100
2	I21	I25	02/01/2100	02/06/2100
3	I21	I25	01/01/2100	02/01/2100
…	…	…	…	…

(c)

Figure 3.13 (a) An event log with recurrent event; (b) an event log after the recurrent event was removed; (c) the pair log after transformation.

The pair log was transformed back into an event log following the Algorithm-2 the reverse version of the Algorithm-1. The final transformation produced an event log for the next stage.

3.2.4 Stage-4: Mining & analysis

In this stage, the event log from the previous stage was loaded into process mining tools for mining and analysis of disease trajectory models. Disease trajectory model was discovered using algorithms provided in the process mining tools. This current work used process mining tools of ProM, DISCO, and Celonis. There are multiple discovery algorithms in ProM that available as plugins, while DISCO and Celonis only provide one type of algorithm – a proprietary algorithm that is based on the Fuzzy Miner [32]. The algorithm mainly used in this current work was the Heuristics Miner

that available as a plugin named *interactive Data-aware Heuristics Miner* (iDHM) in ProM 6.8. The other two process mining tools DISCO and Celonis were used to show how the disease trajectory analysis can be performed using the commercially available process mining tools.

3.2.4.1 Mining disease trajectory model

There is a range of formats to visualise the disease trajectory, starting from a Petri Net, a Heuristics Net, a Fuzzy Model, or as a directly-follows graph.

The event log created from the synthetic data set was loaded into ProM and DISCO for model discovery while ProM was also used for conformance checking. Figure 3.14 shows a disease trajectory model of the feasibility study using DISCO. The model contains nodes and arrows representing the diagnostic codes and the paths, respectively. The model also has a 'start' node represented by a green circle at the top and an 'end' node – the red circle at the bottom of the figure. Each node contains the ICD-10 diagnostic code and a number to show the frequency of the case. The arrows are also numbered to show how many cases that have the sequence. The nodes have different colour intensities while the arrows have different thicknesses. The darker the colour of the nodes meant the more cases had happened, also the thicker the arrows meant there were more cases had happened. The properties to show the increased number of events and cases are also available in iDHM's DFG, and Celonis's DFG.

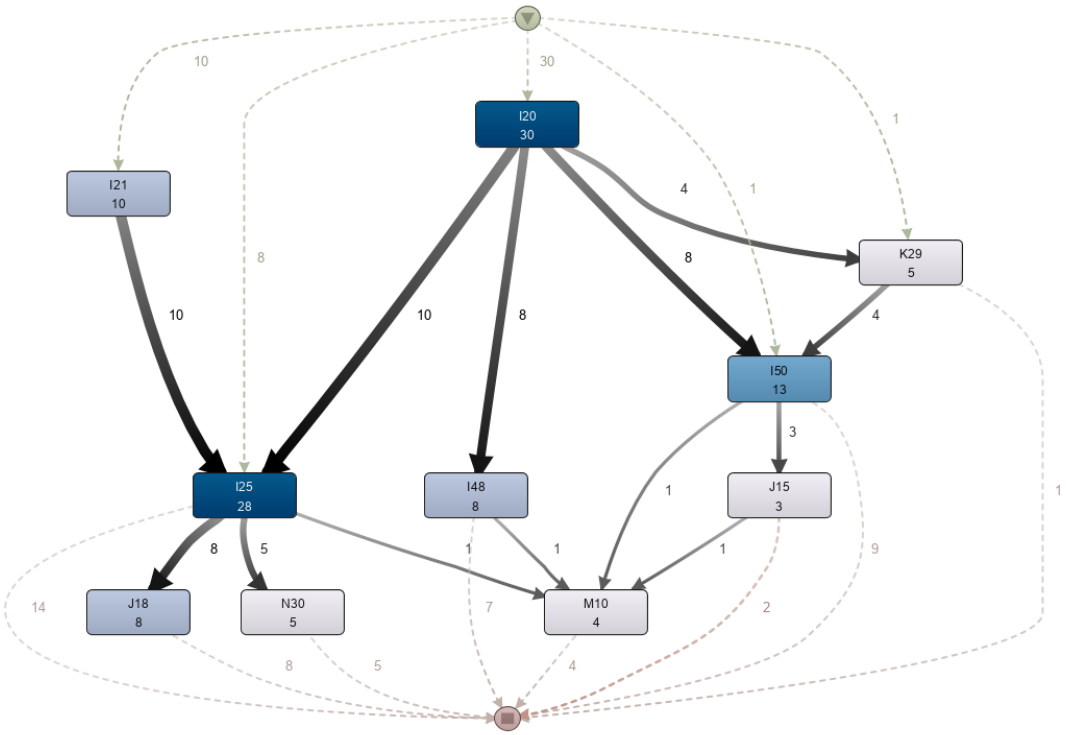
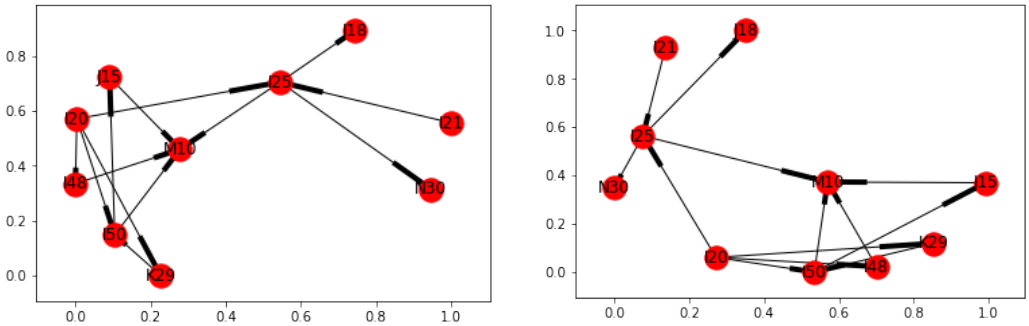


Figure 3.14 Disease trajectory model using the synthetic data set (adapted from [27]).

3.2.4.2 Disease trajectory model analysis

Using the discovered disease trajectory model in Figure 3.14, a visually manual checking was conducted to check if the traces in the Jensen et al.'s model (Figure 3.9) are available. An identifier of pair ID was assigned to each trace in Jensen et al.' model using numbers. The same trace was searched in the discovered disease trajectory model and the same pair ID, and a check sign (✓) was given if the trace was found. Based on the manual checking, all traces available in the Jensen et al.'s model are present in the discovered disease trajectory model.

A Python library named NetworkX was used to make the checking automated. The library implements an isomorphism checker following Cordella et al. (2001) [121] resulting the two models as isomorphic. Figure 3.15 shows the result of the NetworkX Python library.



(a) (b)

```
In [12]: # Check whether the graphs are isomorphic

         nx.is_isomorphic(g_Jensen, g_promin)

Out[12]: True
```

(c)

Figure 3.15 Isomorphism checker using NetworkX Python library. (a) a modelled graph of Jensen et al.'s model. (b) a modelled graph of process mining discovered model. (c) a code snippet showing the result from Jupyter Notebook, "True" means the graphs are isomorphic.

3.2.4.3 Conformance checking

Conformance checking was conducted to get the measurement values of fitness, precision, and generalisation. The plugins mentioned in section 3.1.4.2 were used to measure the quality of the discovered disease trajectory model, and the result is presented in Table 3-4.

Table 3-4 Conformance checking result of the feasibility study.

Fitness	Precision	Generalisation
0.961	0.79	0.963

The maximum score of each conformance checking measurement is 1. Looking at the obtained result above, all three conformance values are high, meaning that the discovered disease trajectory model is representative enough of the event log.

The conformance checking in this current work is to show that process mining has the capability of measuring the quality of the discovered disease trajectory model. In contrast to the already published studies by other researchers using non-process mining approaches, at the time of this thesis being written, no study reported the

quality of the disease trajectory model, whether the model is representative enough of the data.

3.2.5 Stage-5: Evaluation

The evaluation stage was done through several discussions with an expert in the health care domain. This feasibility study was motivated to identify disease trajectories using process mining techniques. Disease trajectory models and process models are identified based on a collection of event data despite they have different contexts. The disease trajectory model was generated from a series of diagnostic codes occurrences while process model was generated from a series of actual activities. Using a simulated data set representing a published disease trajectory, process mining was shown feasible to identify a disease trajectory model using a similar data to the EHRs. The discovered model is qualitatively similar to the published disease trajectory from Jensen et al. (2014). The conformance check results showed that the values of fitness, precision, and generalisation are high so that a conclusion can be drawn where the discovered model is representative of the data.

The process mining approach for mining disease trajectories is suggested as an improvement on the method developed by Jensen et al. (2014) for the following reasons:

1. Process mining approach uses actual traces end-to-end as recorded in the event log to be discovered to produce the model. This approach contrasts with the Jensen et al.'s method where long trajectories were created based on combining direct disease pairs and left out the validation if such long trajectories created are actually available in the data.
2. Process mining method has information that are not available in the method developed by Jensen et al. The durations of the trajectories, the counts of cases, and the quantification of trace variants are provided by some process mining algorithms.
3. Process mining's conformance checking is useful for measuring the quality of the discovered disease trajectory model. In contrast to Jensen et al.'s method, the quality of the discovered model is arguably unknown despite its usage in other disease trajectory analyses [5, 122, 123].
4. Using the process mining method means the model is produced automatically using the various visualisations available. Further forms of support are the

open-source and commercial process mining tools, specialised literature based on health care [59], and a research community that grows worldwide [124].

A representational bias was taken as the limitation of this feasibility study, where the main event is the primary diagnosis that occurred for the first time. This bias is unavoidable in terms of studying model discovery. There are opportunities to include relapsed diseases as recurrent events but at this very particular stage, demonstrating the feasibility of process mining to identify disease trajectory had been achieved.

3.3 Case studies

3.3.1 Case study-1: Experiment using the MIMIC-III data set

The MIMIC-III data set is a database containing a comprehensive information of patients admitted to critical care units in a private hospital named Beth Israel Deaconess Medical Center (BIDMC), Boston, USA [14]. The database contains detailed patient health records of 46,520 individuals during their hospital stay between 2001 and 2012, including patient demographics, diagnostic codes in ICD-9 coding standard, medications, laboratory tests, bedside monitoring, clinicians' notes and reports, and death records. For patients who died outside the hospital, the data were made available by linking to the Social Security Death Index.

The data set is made available online on PhysioNet (https://mimic.physionet.org) by the MIT Laboratory for Computational Physiology and its collaborators [125]. Deidentification was performed before the data were included in MIMIC-III. The deidentification followed the Health Insurance Portability and Accountability Act (HIPAA) standards including removal of eighteen identifiable information such as name, address, telephone number, and dates. The dates were shifted into the future dates using random offset, so they occur between 2100 and 2200 whilst maintaining the time of the day, day of the week and the seasons. The patients of 89 or older were labelled as age over 300 years to obscure their true age. This approach is also suggested with the HIPAA standards.

Knowing that MIMIC-III data set has been deidentified, it brings some challenges to the analysis. Comparison between patients is not possible since the actual time-window of 2001 until 2012 (11 years of duration) were spread into the time-window of 2100 until 2200 (100 years of duration). If there were two patients admitted at the same date and time in the actual data, it is unlikely they still have the same admission date and time after shifting. The sequence of diagnostic codes over time is preserved,

also the time interval between each occurrence, therefore disease trajectory mining will not be affected. There are earlier versions of MIMIC-III data set and this study used MIMIC-III version 1.4 which was released on 2 September 2016.

3.3.2 Case study-2: Experiment using the HES data set

Hospital Episode Statistics (HES) database contains detailed patient-level data on admission, A&E attendances, and outpatient appointments at NHS hospitals in England. The admission data are stored under the Admitted Patient Care (APC) data set within the HES database. The data set used in this study is the HES-APC data set. HES-APC data set has a universal coverage, a long duration of data collection, and contains history of individual patients over time. By having such features, the HES-APC data set is frequently used for research activities [126]. This study used the diagnostic codes of discharge episodes. A discharge episode is the episode to mark the end of a spell, a continuous patient stays in a hospital containing one or more episodes under one or more clinicians. The treating clinician at discharge submits a discharge summary including the diagnoses made and procedures conducted. The discharge summary is forwarded to a clinical coding department to be translated and inputted into a local electronic health records database by clinical coders.

The data stored in local electronic health records are extracted monthly to the Secondary User Service (SUS) – a national data warehouse situated in NHS Digital [127]. Two extractions are needed for the purpose of hospital reimbursement and a provisional monthly HES extract. NHS Digital performs basic data evaluation, cleaning, adding geographical data using patients' postcodes, and adding pseudonymised patient identifiers called HESIDs [128, 129]. Hospitals are then given a provisional annual HES extract by the NHS Digital for a final review and a permission to submit a single further data submission to HES (the 'annual refresh') at the end of the financial year. After checking the annual refresh, NHS Digital makes the final annual HES available for further purposes including research [127].

For this study, the access to the HES-APC data set was granted by the Health & Social Care Information Care (HSCIC) to the Cardiovascular Epidemiology research group at the University of Leeds under the Data Sharing Agreement (DSA) number DARS-NIC-17649-G0X4B-v0.6. The data set contains eight annual HES-APC's (2008/2009 to 2015/2016) until the last available data of 2016/2017 period.

The research was funded by the British Heart Foundation (BHF) under the project grant number PG/13/81/30474. The data set was stored on the University of Leeds's Secure Electronic Environment for Data system and the access was strictly limited.

3.4 Summary

This chapter has explained the general methodology, the data sets, and the feasibility study as the key steppingstone to the rest of this study. The PM^2 was followed throughout the study consisting of five stages: (1) Planning, (2) data extraction, (3) data processing, including the proposed method of event log transformation and pairwise filtering; (4) mining and analysis, and (5) evaluation.

The next two chapters present the two case studies using two different data sets as described in section 3.3. Chapter 4 describes data analysis of the MIMIC-III data as the first case study. Chapter 5 describes the data analysis HES-APC data as the second case study with three different focuses of analyses.

Chapter 4
Case Study-1: Experiments using the MIMIC-III data set

This chapter presents the first case study using the MIMIC-III data set. The data quality assessment of MIMIC-III for process mining was done and published by other members of the research group. The work in this chapter was presented in a virtual international conference on process mining, managed by the University of Padova, Italy, and published in the Lecture Notes in Business Information Processing book series [86].

4.1 Data description

The data set used in this case study is patient clinical records in the MIMIC-III data set. The MIMIC-III data set has been through a set of de-identification process following the Health Insurance Portability and Accountability Act standards. Those include anonymisation process of shifting all timestamps into the future dates between 2100 and 2200 by a random offset generated for each patient. The shifting of the timestamps does not impact the sequence of the activities and time intervals, but comparisons between patients are not possible. This limitation does not affect the analysis of disease trajectories but limits other process mining approaches. For example, the identification of the disease burden over time would not be reliable.
Among other components, the MIMIC-III database contains diagnostic codes in the ICD-9 coding standard. It was possible to extract event data from the table of ADMISSIONS, PATIENTS, and DIAGNOSES_ICD to generate information on the sequence of patients' diagnoses during their admissions. Further description of the experiment of disease trajectory mining using the MIMIC-III data set is described in the following section.

4.2 Experiment

The steps to identify disease trajectories using process mining were mainly based on the Process Mining Project Methodology (PM^2) [10]. The framework consists of six stages: (1) Planning, (2) Extraction, (3) Data Processing, (4) Mining and analysis, (5) Evaluation, and (6) Process improvement and support. The last stage of process improvement and support was not applicable and was not implemented in this experiment. The following subsections describe each of those stages as they were

implemented in the experiment of disease trajectory mining with the MIMIC-III data set.

The PM^2 framework is designed for process mining projects to improve process performance or compliance with rules and regulations. We used the PM^2 for disease trajectory mining with the following adaptations.

- Stage-1 – Planning. The planning was done by identifying research questions from a literature review and confirming those questions by the project team consisting of a clinician, an epidemiologist, and computer scientists.
- Stage-2 – Extractions. The extractions were done by defining the scope based on the granularity level of data, the time and selection of the attributes of interest. For the purposes of capturing the progression of diseases, only those patients with a minimum of two hospital admissions were selected for analysis. The follow-up for the mortality status of these patients was conducted until their last recorded discharge from the hospital, which served as the last censoring date for those who passed away during their hospital stay. In cases where patients died outside of the hospital, the censoring date was recorded in the social security master death index in the MIMIC-III database. In terms of diagnoses, the first three digits of the ICD-9 codes were utilized, while codes not related to disease development, such as administration codes, were excluded.
- Stage-3 – Data processing. The data processing was done by creating an event log as a result of creating views, filtering the log, and enriching it. The case identifier for each event was taken from the patient identifier “subject_id”, the diagnostic code was used as the event name “diagnosis_code”, and the admission time “admittime” as the timestamp. The event log was filtered by removing patient records that only have one diagnostic code recurring diagnostic codes to retain the first occurrence, then reapplying the exclusion of patients who had one diagnostic code.
- Stage-4 – Mining and analysis. This stage was done in ProM to identify unique trace variants, perform process discovery, visualise the discovered model, and check for conformance. Additional process analysis was done to calculate descriptive summary statistics of the identified disease trajectories, including stratification based on patient groups.

4.2.1 Stage-1: Planning

The planning stage aims to setup the analysis by identifying research questions, selecting business processes, and composing project team. The aim of this current experiment is to mine disease trajectories from MIMIC-III data set – an actual electronic health record that came from a private hospital in Boston, USA. The scenario of using this data set is by taking all the available data without any specific selection of diseases and time windows.

We defined the main research questions from a literature review confirmed by clinical experts. The main research questions of this experiment are:

Q1. Can disease trajectories be identified using a process mining approach?

Q2. What are the most followed trajectories and how long were the duration trajectories?

Q3. Are there differences in trajectories followed by different patient groups (based on sex, age group, and mortality status)?

Q4. What are the longest and shortest average time transition trajectories?

We composed a project team consisting of epidemiologists, health-science methodologists, clinical experts, and computer scientists. The decision to include the experts was suggested in the PM^2 framework where in this current experiment a clinical expert was involved. During all stages of the project, all experts in our team contributed to execute and evaluate the results and suggest strategies for the next stages.

4.2.2 Stage-2: Extraction

The extraction activity in the second stage was started with the complete MIMIC-III data set consisting of 58,976 unique admissions from 46,520 patients. Of those admissions, there were 6,984 unique ICD-9 diagnostic codes used to identify 651,000 diagnoses. From this data set, there were 172,685 (26.5%) diagnostic codes were excluded because they are medically known to be related to external factors not directly related to the development of diseases [3], these including pregnancy (ICD-9 3-digit codes 630-679, 760-779), general symptoms and signs not related to a disease (780-799), external cause (800-999, E800-E999), and administration (V01-V89).

Following the diagnostic process in every patient encounter with the hospital, the clinician records the main cause of the patient's hospitalisation, including any

following conditions. The recorded diagnostic code of the main reason is given the highest priority as the primary diagnostic code. The rest of the codes are considered secondary diagnostic codes. A number is assigned sequentially to each diagnostic code to determine the priority. These numbers are stored in the SEQ_NUM variable. The primary diagnostic code is given by SEQ_NUM 1, followed by SEQ_NUM 2, 3, 4 and so on for the secondary diagnostic codes. The information on the priority enables the clinician to design the intervention regime to help the patient. There were 436,483 (67%) secondary diagnostic codes excluded to focus on the 41,832 primary diagnostic codes. The final data set to be used in this experiment consists of 4,911 patients with 11,725 admissions and 350 diagnostic codes. The data is also segregated by gender to understand the different behaviour of disease progression between male and female patients. It has been known that the progression of disease and outcome are affected by genetic factors including the presence of chromosomes X and Y [130, 131]. Table 4-1 and Figure 4.1 show the patient characteristics and the 20 most common diagnostic codes, respectively.

The data is also segregated by gender to understand the different behaviour of disease progression between male and female patients. There are significant evidences that the disease progression and outcome is genders biased [126, 127].

Table 4-1 Patient characteristics of the final data set for experiment.

Characteristics	**Patients (N = 4,911)** ***no. (%)***
Age group	
14-17	1 (0.02)
18-24	42 (0.86)
25-34	124 (2.52)
35-44	329 (6.70)
45-54	721 (14.68)
55-64	1,032 (21.01)
65-74	1,168 (23.78)
75-84	1,046 (21.30)
85-89	170 (3.46)
>89	278 (5.66)
Sex	
Female	2,174 (44.27)
Male	2,737 (55.73)
Mortality status	
0 (Censored)	3,880 (79.01)
1 (Dead)	1,031 (20.99)

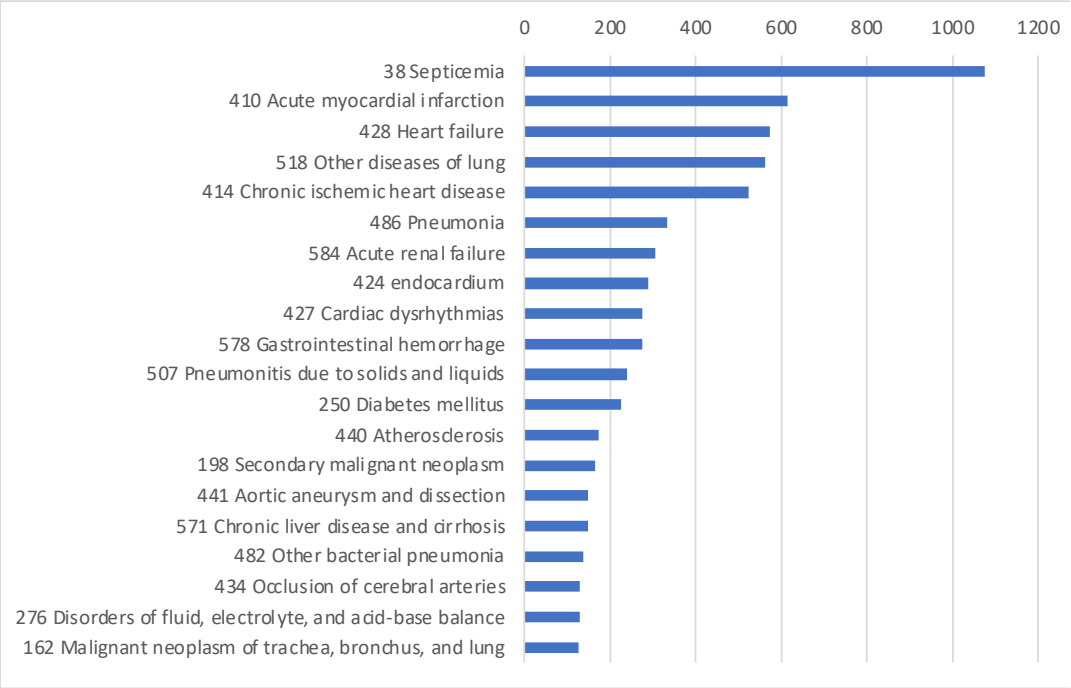


Figure 4.1 The 20 most common diagnostic codes.

4.2.3 Stage-3: Data processing

In Stage-3 – Data processing, an event log was created from the data set. The case identifier for each event was taken from the patient identifier (subject_id), the diagnostic code was used as the event name (diagnosis_code), and the admission time as the timestamp (admittime). Table 4-2 presents the tables and their respective field names of MIMIC-III database as the data source of the current experiment.

Table 4-2. The mapping of MIMIC-III's tables and field names to create event log.

Variables	Table name	Field name
Case identifiers	Patients	subject_id
Event	Diagnoses_ICD	Hadm_id, icd9_code, seq_num
Activity name	Diagnoses_ICD, Admissions, Patients	Icd9_code (first 3 digits) Hospital expire-flag Expire_flag (translated into 1: Dead, 0: End of data)
Time stamps	Admissions, Patients	Admittime, dischtime, deathtimes Dod, dod_hosp, dod_ssn
Sex	Patients	Gender
Age*	Patients, Admissions	Dob, admittime
Age group**	Patients, Admissions	

* The age calculation using patients' DOB and Admissions' admittime.

** the variable was added to group the patients' age.

Variable selections were done to meet the minimum requirements of an event log. There were 2,692 (16.2%) recurrent diagnoses that were removed, retained the first occurrence of the diagnoses only. Patients with only one admission were excluded. The final event log consists of 4,911 patients with a total of 11,725 admissions and 350 diagnostic codes. Each subject_id in this final event log contains a temporally informed sequence of diagnostic codes. The illustration of the event log and the trace of diagnosis are presented in Figure 4.2(a) and Figure 4.2(b), respectively. This final event log was then transformed into pair log (see Figure 4.2(c)).

The transformation of event log into pair log consisted of 6,814 ordered pairs of diagnostic codes (D1→D2). After filtering for RR > 1 and the binomial tests for directionality there were 3,781 pairs remaining. It was also suggested that there were 826 ordered pairs of diagnostic codes with a statistically significant dominant direction.

The 826 ordered pairs were then used as a reference to make a selection of the paired diagnostic codes from the pair log. The final data set contained 796 patients with 3,218 admissions and 49 diagnostic codes.

subject_id	diagnostic_code	timestamp
21	410	11/09/2134 12:17
21	038	30/01/2135 20:50
124	433	24/06/2160 21:25
124	441	17/12/2161 03:39
124	440	21/05/2165 21:02
124	569	31/12/2165 18:55

(a) The extracted event log

#21: 410→038

#124: 433→441→440→569

(b) The trace of diagnosis

subject_id	Antecedent	Subsequent	Time1	Time2
21	410	038	11/09/2134 12:17	30/01/2135 20:50
124	433	441	24/06/2160 21:25	17/12/2161 03:39
124	441	440	17/12/2161 03:39	21/05/2165 21:02
124	440	569	21/05/2165 21:02	31/12/2165 18:55

(c) The *pairlog*

Figure 4.2 The illustration of transforming event log into pair log. (a) Event log extracted from MIMIC-III; (b) traces of diagnostic codes from an event log (a); (c) the result of transforming event log.

4.2.4 Stage-4: Mining and analysis

Stage 4 (Mining and Analysis) was done in ProM. Upon loading the event log into ProM, the event log was analysed to identify the trace variants. Moreover, 20 traces were excluded and then visualised using the Explore Event Log (Trace variants/ Searchable/ Sortable) [132]. The process discovery was done using the Interactive Data-aware Heuristics Miner (iDHM) to discover the disease process models [115].
The conformance checking was done using replay fitness, precision, and generalisation metrics [36]. Replay fitness measures how many traces from the log can be reproduced in the process model, with penalties for skippings and insertions. Precision measures the model if it is underfitting, i.e., the model limits various behaviour that is different from what is discovered in the log. Generalisation measures the redundancy of nodes in the model, with more redundancy means more variety of possible traces can be represented. Generalisation should not only allow the behaviour in the event log. The values of those three metrics were represented by a number between 0-1, with a higher value representing better conformance. The plugins used in the conformance checking are:

- The Replay a Log on Petri Net for Conformance Analysis [117] for replay fitness,
- The Align-ETConformance [119] for precision,
- The Measure Precision/ Generalization plugin for generalization.

Eighty-one distinct trace variants were used to guide the process discovery algorithms in answering Q1. The results of the discovered disease trajectory model showed a high level of conformance, with values of fitness = 0.9299, precision = 0.9381, and generalization = 0.9177. To further evaluate the model, a 5-fold cross-validation was performed, where the original event log was randomly divided into five sub-event logs of equal size. One sub-event log was used as validation data while the remaining four served as training data, and this process was repeated five times to ensure that each sub-event log was used once as validation data. The average results from the cross-validation were lower than the conformance, with fitness = 0.9195 (SD: 0.006), precision = 0.8221 (SD: 0.06), and generalization = 0.8887 (SD: 0.02). This indicates that the discovered trajectory model, depicted in (Figure 4.3) is robust to sampling, allows for the representation of traces seen in the event log, is precise in excluding behaviour not present in the event log, and is general enough to predict future behaviour of the trajectories.

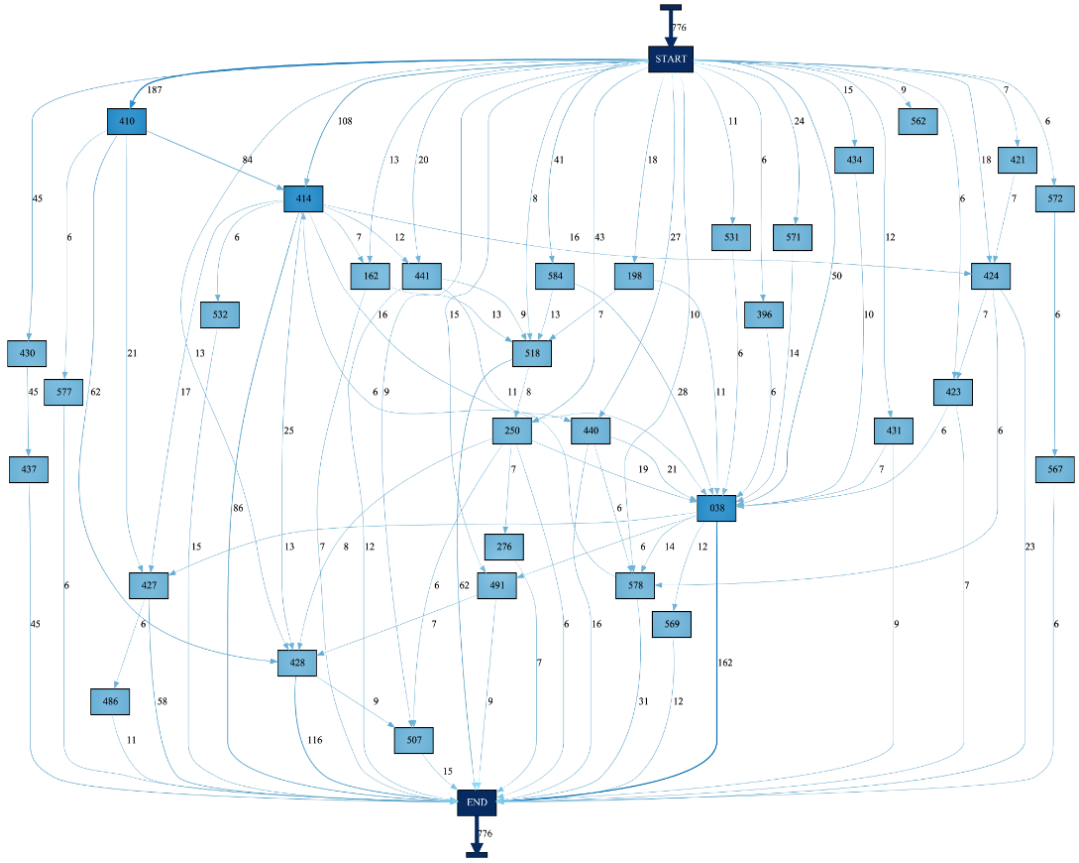


Figure 4.3. The directly-follows graph representation of Disease Trajectory Model of Critical Care patients in MIMIC-III with the minimum case frequency = 6.

The result of Q2 showed that among 776 patients, 81 different disease trajectories were identified (as shown in Table 4-3). The most common trajectory (n=80; 10.3%), was from acute myocardial infarction to ischemic heart disease, which aligns with previous research findings [3, 133, 134]. Septicaemia was the most prevalent condition, with a total of 212 patients (27.3%) experiencing it either before (n=50; 6.4%) or after (n=162; 20.9%) another diagnosis, with 143 patients (66.9%) passing away. This supports previous research that sepsis is linked to both morbidity and mortality [14, 135]. There were also three uncommon trajectories followed by two patients (0.26%) as shown in Table 4-3.

Table 4-3. The three most-common and least-common trace variants.

#	Traces (%)	Trace Variant*	Median (months)	Dead (%)	Male (%)
1	80 (10.31%)	START→410→414→END	6.5	75	70
2	62 (7.99%)	START→410→428→END	3.9	72.58	54.84
3	45 (5.80%)	START→430→437→END	3.9	4.44	35.56
…	…	…	…	…	…
79	2 (0.26%)	START→410→427→486→END	28.3	100	50
80	2 (0.26%)	START→507→491→482→END	43.6	50	100
81	2 (0.26%)	START→518→250→038→END	14.6	100	0

**ICD-9 Codes translation: 038= Septicaemia, 250= Diabetes mellitus, 410= Acute myocardial infarction, 414= Ischemic heart disease, 427= Cardiac dysrhythmias, 428= Heart failure, 430=*

Subarachnoid haemorrhage, 437= Other and ill-defined cerebrovascular disease, 482= Other bacterial pneumonia, 486= Pneumonia, organism unspecified, 491= Chronic bronchitis, 507= Pneumonitis due to solids and liquids, 518= Other diseases of lung.

The third question addressed (Q3) was if there were any differences in disease trajectories followed by different patient groups. The question was answered by comparing the trajectories between male and female patients and between different age groups (18-34 years, 35-64 years, and over 64 years). The male group consisted of 447 patients with a median follow-up duration of 6.98 months (IQR 1.6 – 28.2) and 252 cases (56.3%) ended in death. The most prevalent disease trajectory was acute myocardial infarction followed by other forms of chronic ischemic heart disease (56 cases, 12.5%) with a median interval of 6.5 months (IQR 1.5 – 35.3). The female group consisted of 329 patients with a median follow-up duration of 7 months (IQR 2 – 24.4) and 176 cases (54.4%) ended in death. The most prevalent disease trajectory was subarachnoid haemorrhage followed by other and ill-defined cerebrovascular diseases (29 cases, 8.8%) with a median interval of 3.4 months (IQR 2.3 – 7.5).

The results above suggest that the trajectory of AMI → ischaemic heart disease is the most common in male patients. AMI is recognised as being more common in men rather than women. Other contributors are smoking habits, diabetes, and high blood pressure. The result conforms with the previous study [136]. Compared to the female patients, the most common trajectory is subarachnoid haemorrhage → ill-defined cerebrovascular disease. The result confirms the study of Wang et al. [137], where the disease of subarachnoid haemorrhage is common in female patients due to estrogen deficiency that occurs in the late 30s. The reduction of estrogen dramatically contributes to the loss of collagen as the blood vessel protector."

For the 18-34 year old group, the most common disease trajectory was diabetes followed by hypertensive chronic kidney disease with a median time interval of 55.8 months (IQR 33 - 56.5). For the 35-64 year old group, the most frequent trajectory was acute myocardial infarction followed by ischemic heart disease, with a median interval of 7.8 months (IQR 1.9 - 39.7) and 40.4% of cases ending in death. The most common trajectory in the group over 64 years was acute myocardial infarction followed by heart failure, with a median time of 4.7 months (IQR 1.5 - 21.8) and 68.1% of cases ending in death. In terms of the longest and shortest average time transitions, the longest was ischemic heart disease to Diverticula of intestine at 63 months and the shortest was Gastrointestinal haemorrhage to Liver abscess and

sequelae of chronic liver disease, with an average time transition of less than a month (0.98) (Table 4-4).

Table 4-4. The three longest and shortest average time interval trajectories in MIMIC-III

Antecedent	Subsequent	Mean*	Median (IQR)**
A. The three longest average time interval trajectories (descending)			
Chronic ischemic heart disease	Diverticula of intestine	63	75.9 (54 – 84.8)
Chronic ischemic heart disease	Occlusion of cerebral arteries	52.7	51.2 (40.4 – 52.6)
Chronic ischemic heart disease	Heart failure	46	41.5 (4.6 – 89.7)
B. The three shortest average time interval trajectories (ascending)			
Gastrointestinal haemorrhage	Liver abscess and sequelae of chronic liver disease	0.98	0.81 (0.6 – 1.3)
Other diseases of endocardium	Other diseases of pericardium	1	0.8 (0.6 – 1.13)
Chronic bronchitis	Other bacterial pneumonia	2.2	2.2 (1.6 – 2.7)

**Mean and Median were calculated in months, **IQR = interquartile range*

4.2.5 Stage-5: Evaluation

The results of each stage were discussed and evaluated through discussions with clinical experts. The study utilized the MIMIC-III data set which is similar to various electronic health record (EHR) systems utilized in hospitals globally. The PM^2 framework was employed to extract a representative disease trajectory model from the EHR and examine quality dimensions. This research opens the door for future work in applying the technique to larger EHR data sets.

One advantage of using process-mining to develop disease trajectories is that it can present a summary of cases, events, and the time between them. Our process-mining approach identified a trajectory of acute kidney injury followed by septicaemia, with an average time of 16.22 months, which is in line with previous studies [138] that have found sepsis to be a frequent result of AKI in intensive care. The process-mining method can also estimate the likelihood of sepsis developing after AKI [139]. Furthermore, our approach includes additional patient attributes that make it possible to categorize outputs based on characteristics such as sex, age group, and mortality status. For instance, the process-mining tools allowed us to identify a dominant trajectory among females as subarachnoid haemorrhage (ICD-9 code: 430) followed by other and ill-defined cerebrovascular disease (ICD-9 code: 437) – as suggested in the previous research [140].

4.2.6 Pareto Principle's Composition Ratio

Pareto Principle is used for data filtering at the beginning of the disease trajectory analysis. The principle states that 80% of outcomes come from 20% of causes. Nevertheless, the composition ratio is sometimes flexible, no need to follow the 80/20 fashion [107]. The principle applies by filtering diagnostic codes that cause 80% of the most common causes of hospitalisation. The selected diagnostic codes are used to create pair of diagnostic codes and identify the significantly dominant disease trajectories. This current work presents how the changes in the Pareto Principle's composition ratio affect the number of significantly dominant pairs of diagnostic codes.

The event log extracted from the MIMIC-III data set (as used in Case Study-1) was filtered using various compositions of the Pareto Principle. The filtered event log was transformed into a pair log and recorded the number of pairs. Figure **4.4** presents the number of diagnostic pairs and the significant pairs following the Pareto Principle's composition ratio changes.

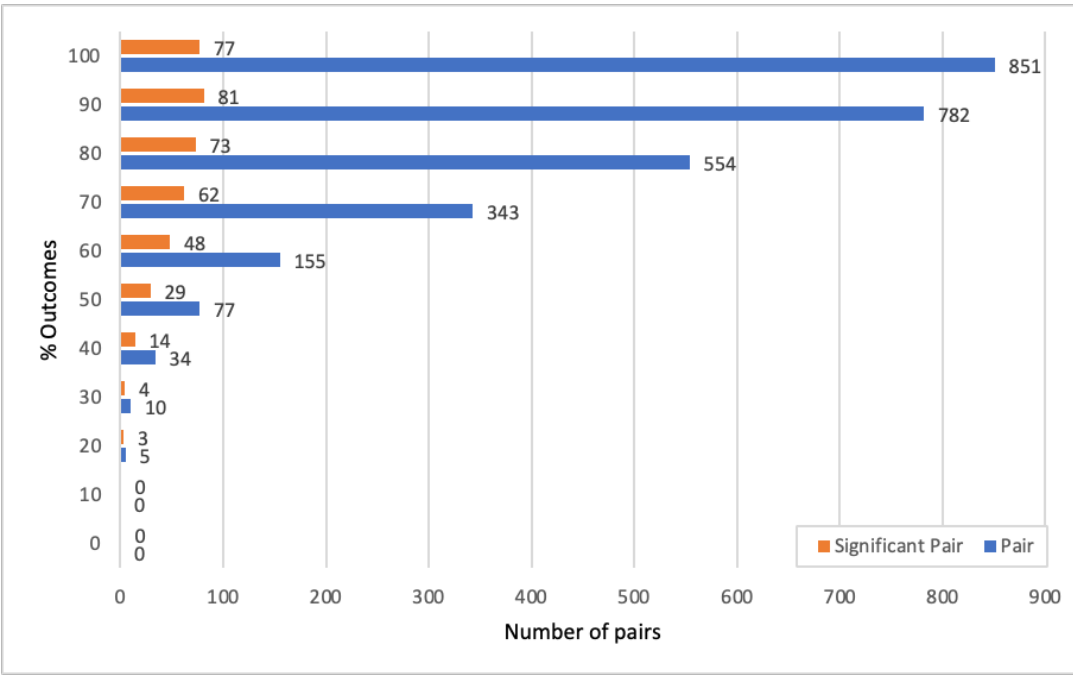


Figure 4.4. Number of pairs and significant pairs following the Pareto Principle's composition ratio.

The diagnostic codes composition ratio of 0-100, 10-90, 20-80 until 100-0 produced various number pairs and significant pairs. Both numbers are increasing following the composition ratio changes. Unexpectedly, the highest number of significant pairs

occurred when the composition ratio was 90-10 and then dropped at the composition of 100-0. The figure shows that the changes in the Pareto Principle's composition ratio biased the number of the significant pair. Despite this variation, the Pareto Principle remains useful for identifying and prioritising the most important factor in a given situation.

4.3 Future work

The current work can be upscaled by implementing the method using a wider coverage of data such as a population-wide EHR. Using a wider coverage of data would allow for capturing patients' admission to any hospitals within a certain boundary e.g., national, or regional, instead of using data from one hospital. The mobility of patients allows them to get admitted to any nearest hospital. One of the challenges of using such data set is the availability of a single patient identifier that is recognisable and shareable by any different health care organisations within the country. Another opportunity is to perform a comparison study to see the similarities or dissimilarities of the progression of disease between hospitals.

In terms of the method, the association of diagnostic pairs analysis can be improved using a null hypothesis significance testing. The current work only relies on the measure of RR > 1, by applying the significance testing will increase the validity and the reliability of the outcomes.

4.4 Summary

The work in this chapter has contributed to the community of health care and process mining, especially the Process Oriented Data Science for Health care. The implementation of process mining techniques to identify disease trajectory from an actual EHR was implemented. Event log transformation into pair log was required to allow the statistical analyses: measuring association between pairs of diagnostic codes and testing the directionality of the pairs to produce the trajectory. The results were discussed and validated by the statistical experts and clinical experts. The common disease trajectories mined from the MIMIC-III data set were supported or in line with other medical publications.

This finding suggests that the method of identifying and producing visualisation for disease trajectory analysis using a range of standard process mining tools became of

interest to health care researchers. Process mining was able to identify the Septicaemia as a common problem in a critical unit environment including the duration of the occurrence. This finding may open an opportunity for further exploration in prediction or prevention to improve the quality of health care outcomes.

Chapter 5
Case Study-2: Experiments using the HES-APC data set

Chapter 4 presented the work of identifying disease trajectories using process mining approach using a hospital-scale electronic health record without any pre-selection on a specific disease. In contrast to the previous chapter, Chapter 5 presents the second case study – the implementation of the process mining approach to identify disease trajectory models using electronic health records collected from the National Health Service (NHS) hospitals in England named the Hospital Episode Statistics-Admitted Patient Care (HES-APC) data set. This chapter started with the descriptions of the HES-APC data set (Section 0), in terms of its provenance, characterisation, acquisition process and quality. The experiments in this current work are divided into three sub-experiments using a large amount of data from the HES-APC data set. The behaviour of the process mining method using a large data set was studied in this current work. The first experiment (Section 5.2) was on process mining of disease trajectory using five thousand randomly selected patients. The second experiment (Section 5.4) was on process mining of disease trajectory using five thousand randomly stratified patient selection. The third experiment (Section 5.5) was on process mining of disease trajectory using a full set of the HES-APC data extract, and the fourth experiment (Section 5.5) was mining patients' disease trajectories after the event of acute myocardial infarction. A manuscript is under preparation for a journal submission in The Lancet.

5.1 Data description

5.1.1 Data provenance

Hospital Episode Statistics (HES) is a nation-wide database containing detailed records of all inpatients, outpatients, accident and emergency, and adult critical care at NHS hospitals in England. The need to develop this database was started in the 1979 Royal Commission [141]. This commission recommended to establish a Steering Group on Health Services Information, to provide the information needed to support decision makers in the NHS England. The steering group enforced that *"improved data would help to improve the quality and efficiency of the NHS"* [142]. In the early 1980s, the steering group produced a series of six reports providing recommendations on improving data collation, processing, use and governance within

the NHS. The activity of collecting national data was started in 1987 in response to a recommendation by the England's Department of Health working group in the early 1980 [143]. The main purpose of collecting national hospital activity data is to inform management and planning services.

In 1989, the English HES-APC database was established to record every 'episode' of patient treatment in a hospital. HES-APC data covers all NHS Clinical Commissioning Groups (CCGs) in England, including private patients treated in NHS hospitals, patients resident outside of England, and care delivered by treatment centres (including those in the independent sector) funded by the NHS. This database was part of the Commissioning Data Set (CDS), containing data collected during patients' time locally at hospitals. Those data were submitted monthly to the NHS Digital. The NHS Digital consolidated, validated, cleaned, and returned it to health care providers as the Secondary Uses Services (SUS) data set. The SUS data set includes information needed for reimbursement for treatments undertaken by patients [144].

The HES database can also be used for secondary purposes, including research and health service planning. The HES can be used to support a wide range of research purposes, such as to monitor trends and patterns in NHS hospital activity, assess effective deliver of care, and reveal health trends over time. HES database has been used in standalone studies and also has been linked to registers and other sources of information, linked into longitudinal observational studies, linked to randomised controlled trial samples for log-term outcome assessments, and also linked to research repositories, such as the Clinical Practice Research Datalink [145, 146].

Each record in the HES contains clinical information of an individual patient admitted to an NHS hospital, including:

1. Clinical information about diagnoses and operations,
2. Patient information, such as age group, gender, and ethnicity,
3. Administrative information, such as dates and methods of admission and discharge,
4. Geographical information such as the location of patient treatment and the location of a patient residential.

NHS Digital runs the Spine service for the NHS through the Digital Delivery Centre. The Spine is a central, secure system for patient data in England. Recent advancements include enhancing the security of child information through the Child

Protection Information System and creating the Spine Mini Service for simplified access to demographic data [147]. Users primarily access Spine through clinical systems or through the Spine portal. The Spine enables secure sharing of information through nationwide services such as the Electronic Prescriptions Service (EPS), Summary Care Record, and the e-Referral Service.

The data were cleaned by the NHS by following the auto cleansing guideline [148] to improve the consistency and usability of HES data. These rules are used to clean common and obvious data quality errors and derive additional data items to populate the HES data set. The rules that apply to dates will check that the data is not before the start of the data year (year start) or after the end of the period being processed (period end). The data year starts every 1 April and ends on 31 March a year after. In most cases, each rule cleans only one data item, but there are a few instances where a rule cleans or derives a set of data items. An example is the postcode-derived items [148].

The HES database is a comprehensive collection of information on inpatient hospital stays, outpatient visits, emergency department visits, and adult critical care in NHS hospitals in England. It also includes records for patients who paid for their own treatment by an NHS provider, records for non-English residents, and records for treatments funded by the NHS but provided by independent (non-NHS) providers. [149]. The HES contains every ‘episode’ of admitted patient care, counted by completing care with a consultant. It means that more than one single stay in the hospital can be recorded as more than one episode in the database. In the NHS in England, this involves tracking around 16 million ‘episodes’ of care every year [144]. The HES patient ID (HES ID) makes it possible to track patients through the HES database. This HES patient ID is a central concept of many HES outputs, including spell construction, emergency readmission and linkage to other data sets, such as mortality.

From April 2021, there were some changes to the processing of the HES data set to provide data users with the tools to process larger volumes of data and manage the accuracy, usefulness, and security of the incoming data. The changes include: (1) the introduction of the Master Person Service (MPS) Person ID to enable direct linkage of patient records across data sets, reducing the time and complexity in the HES data analysis, (2) the HES ID and some other invalid fields will be retired, and (3) reference data will reflect real-time changes in organisations. Those changes will help in several

areas, such as improving patient matching across all national patient-level data sets; improving data quality, coverage and timeliness; allowing self-service; supporting new analytical and data science capabilities, and increasing efficiency through automation, among others [150]. Although those changes do not apply to the data set used in this thesis, it is important to note that the data set in this thesis is the previous version of the HES data.

5.1.2 Data characterisation

5.1.2.1 Episode, spell, and discharge

The record contains episodic information about patient care during his/ her stay in a hospital. One episode represents the duration of a patient encounter with one consultant for care [151]. A spell, or admission, is the duration a patient stays within a hospital. During the stay, a patient may encounter more than one consultant, meaning a patient may have more than one episode. Every finished care from one consultant is called a Finished Consultant Episode (FCE). If a patient is discharged after one FCE, the patient only has one episode spell or one episode admission. The illustration below represents a record of a patient who encounters multiple consultants until discharge.

For example, Patient A was diagnosed with a heart attack and treated by Consultant A (Episode-1). Patient A was also diagnosed and treated for diabetes by Consultant B (Episode-2) during his/ her hospital stay; after several episodes, Patient A was discharged by Consultant XYZ. The episode and the spell are finished when a patient is discharged, dies, or transferred to another health care provider.

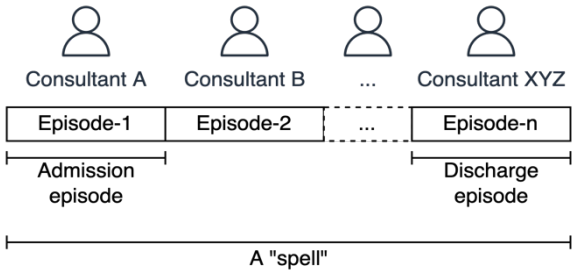


Figure 5.1 Illustration of a spell, episodes, consultants, admission, and discharge (adapted with modification from [151]).

If the patient is transferred into a different hospital, then the episode and the spell in the current hospital are ended. A new episode and a new spell will be created at the new hospital. Figure 5.1 illustrates the given example.

Each patient has a unique identifier named HESID, a unique identifier EPIKEY is assigned for each episode. In this study, the HESID is anonymised and stored under the field of ENCRYPTED_HESID. This field contains a combination of alphanumeric characters with a length of 32.

HES-APC data is structured to follow the duration of a financial year. One financial year started on 1st April and ended on 31st March of the following year. A spell may start or end at a different financial year and there are two indicators in the data set 'SPELBGIN' and 'SPELEND' to flag if an episode is the beginning or the end of a spell. The new clinician at the new hospital who is responsible for Patient A requires to finalise a discharge summary upon the discharge. The discharge summary contains the diagnoses and procedures that were conducted to the patient.

Figure 5.2 shows the relationship of episodes and spells across financial years. Spells are shown as boxes surrounding smaller boxes representing episodes. Assume the current financial year is 2009/2010, then spell A – a three episodes spell, it has two episodes that finished under the current financial year while the last episode is not finished in the current financial year. Spell B only contains one episode that finished under the current financial year; this representation also works for a day case where one episode is started and ended on the same day. Spell C contains multiple episodes that spread across three financial years, it started in 2008/2009 and is still not finished in the current financial year. Spell D has three episodes where the first episode had finished in the previous financial year (2008/2009) and the last two episodes were finished in the current financial year. Spell E contains two episodes where both are finished in the current financial year.

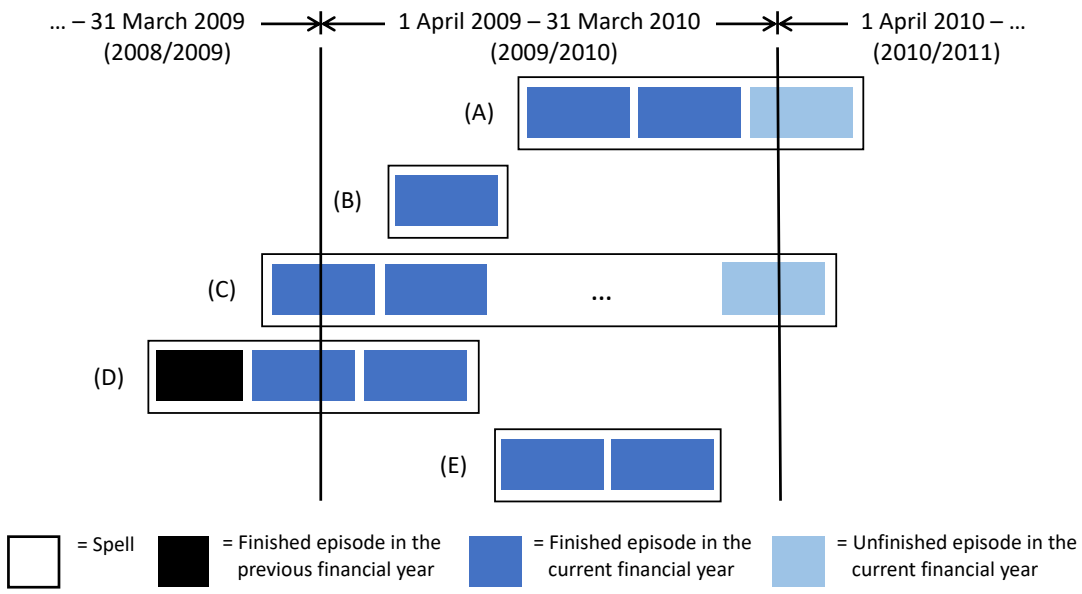


Figure 5.2 Illustration of spells, episodes, and financial years. Figure adapted from Figure 2 in [151].

5.1.2.2 Data selection

The patients' disease journey involves many variables to be recorded in EHR. In the HES-APC data set, the variables are stored as columns containing data. There were 119 variables identified from the requested extraction of the HES-APC data set, covering the patient identity, social data, medical events, diagnoses, and operations. There were 18 variables selected for this current work as candidates to be included in an event log or to facilitate the creation of an event log. Decisions of the variable selection to construct an event log are described below.

1. Case ID

   There were two variables that contain identifier data: ENCRYPTED_HESID and EPIKEY. ENCRYPTED_HESID, is a pseudonymised patient identifier from the original HES identifier (HESID) to increase the patients' anonymity. HESID data are passed between hospitals and other health care providers if transferring patients is required.

   EPIKEY is a record identifier created by the HES system. EPIKEY is an episode identifier that is created every time a patient makes an encounter with a health care provider to receive treatment. A single value of ENCRYPTED_HESID represents a single individual that allows to have one or more EPIKEYs. On the other side, a disease trajectory contains a *portfolio* of diseases of an individual during their lifetime, therefore, an identifier to

represent an individual is needed. The suitable identifier to represent a case in this current work is ENCRYPTED_HESID.

2. Activity

   The HES-APC data set contains twenty variables to record one primary diagnostic code and up to 19 secondary diagnostic codes. The column names are DIAG_01, DIAG_02, DIAG_03, …, DIAG_20. This current work was limited by only selecting the primary diagnostic codes, meaning only the column of DIAG_01 was taken as the activity.

3. Timestamp

   There were 34 variables containing timestamp data, but the level of details was limited. Only months and years were provided as the lowest abstraction of the timestamp data. The removal was done by the NHS Digital to avoid re-identification. Among the 34 variables, there were 24 variables available to record operation dates. The other ten variables were admission date (MYADMIDATE), discharge date (DISDATE), episode end date (MYEPIEND), episode start date (MYEPISTART), financial year (FYEAR), date of birth (MYDOB), PARTYEAR, the status of episode as the beginning of a spell (SPELBGIN), the status of episode as the end of a spell (SPELEND), and submission date (SUBDATE). The selected variables for this current work were MYADMIDATE, MYEPISTART, MYEPIEND, and DISDATE.

4. Resource

   Additional information regarding the patients is stored as resource variables in the event log. These variables contain data on patients' gender, mortality status, and date of birth. The related variables were taken from the HES-APC data set by including columns SEX, MORTALITY, and MYDOB, which contains months and year of the patient's date of birth.

5. Other

   Other related variables to facilitate cleaning and selecting data were also included. The selected variables were admission method (ADMIMETH), discharge method (DISMETH), episode status (EPISTAT), and episode order (EPIORDER).

Table 5-1 presents the summary of the selected variables for this current work, and Appendix A presents the data dictionary from NHS Digital for the selected columns.

Table 5-1 Selected variables from the HES-APC data set.

Identifier	Activity	Timestamp	Resource	Other
ENCRYPTED_HESID, EPIKEY	DIAG_01	MYADMIDATE, MYEPISTART, MYEPIEND, and DISDATE	SEX, MORTALITY, and MYDOB	ADMIMETH, DISDEST, DISMETH, EPIORDER, EPISTAT, EPITYPE, SPELBGIN, and SPELEND

5.1.3 Data acquisition process

Extract of the HES-APC data set was requested to the NHS Digital by the Cardiovascular Epidemiology Research Group within the Leeds Institute of Cardiovascular and Metabolic Medicine at the University of Leeds. The request was granted under Data Sharing Agreement (DSA) number DARS-NIC-17649-G0X4B-v0.6 and the data set was received in nine separate text files (.txt), each file contains data from one financial year. **Table 5-2** shows the detail information on the nine files received from the NHS Digital.

Table 5-2 Summary of the HES-APC data set received from the NHS Digital.

#	File name	Annual reporting period	# Of–Row	File size
1	NIC17649_APC_0809.txt	01/04/08 - 31/03/09	13,0–9,872	9.89 GB
2	NIC17649_APC_0910.txt	01/04/09 - 31/03/10	15,03–,887	11.48 GB
3	NIC17649_APC_1011.txt	01/04/10 - 31/03/11	15,83–,515	12.62 GB
4	NIC17649_APC_1112.txt	01/04/11 - 31/03/12	16,24–,593	12.95 GB
5	NIC17649_APC_1213.txt	01/04/12 - 31/03/13	16,61–,561	13.20 GB
6	NIC17649_APC_1314.txt	01/04/13 - 31/03/14	17,17–,190	13.63 GB
7	NIC17649_APC_1415.txt	01/04/14 - 31/03/15	17,83–,991	14.13 GB
8	NIC17649_APC_1516.txt	01/04/15 - 31/03/16	18,43–,662	14.60 GB
9	NIC17649_APC_1617.txt	01/04/16 - until the latest available data when the request was approved at 13/02/2017.	15,733,889	12.48 GB
	Total		145,913,610	114.98 GB

Upon receiving, the files were loaded in the Secure Electronic Environment for Data (SEED) at the University of Leeds as one table in Microsoft SQL Server under a database named HES_CVEPI. The loading process was done by a member of the

Integrated Research Campus (IRC) Data Service Team at the Leeds Institute of Data Analytics (LIDA).

Access to the HES-APC data set was strictly limited and made available for authorized individuals in the Cardiovascular Epidemiology Research Group. The access was given by joining the research group after completing the Research Data and Confidentiality course by the Medical Research Centre and signing a confidentiality agreement as part of a Secure Electronic Environment for Data Information Governance Policy version 3.0 (last update by 17 February 2014).

The Cardiovascular Epidemiology Research Group requested the HES-APC data set for a specific purpose, to study the pattern of patients' hospitalisation after heart attack or *acute myocardial infarction* (AMI) in medical terms. This current case study contributed to analysing the pattern of disease trajectories to discover the burden of hospitalisation post-AMI.

5.1.4 Data quality

A data quality assessment was done by following the Care Pathway Data Quality Framework (CP-DQF) [152] based on the data quality assessment by Weiskopf & Weng [82]. The framework also suggests that nine entities of the event log [153] be assessed based on four data quality issues to identify the data set [154]. The nine entities of event log are: *case* – the event log contain multiple cases, *event* – an ordered list of event as part of case, *relationship* – there is a relation between event and case, *c_attribute* – attribute of a case, *activity name* – the attribute of an event, *timestamp, position, and resource* – the three optional attributes of an event where timestamp and position are the attribute to order the events in a case, and finally *e_attribute* – attribute of an event.

The four data quality issues to assess the nine event log entities are as follow:

1. Missing data: one or more mandatory properties of process mining are *vacant*.
2. Incorrect data: the information is logged *incorrectly*.
3. Imprecise data: the information is *loss of precision*.
4. Irrelevant data: the *as-is* information may not be relevant.

For each data quality issue, there are three issue indicators: 'N' if the issue is non-existent, 'L' if the issue does exist but less frequent, and 'H' if the issue does exist and high frequent. The above event log entities and data quality issues are combined resulting in up to 36 event log quality issues. Any quality issues that are not applicable

can be left empty. For example, each record in the HES-APC data set is relevant to a certain patient, thus the *irrelevant – case* data quality issue is left empty.

The identification of data quality issues was done carefully and simultaneously with data cleaning and selection where every step was consulted with the clinical expert. This strategy was taken due to the size of the data set and the storage limitation of the database server for log transaction. A STROBE diagram [155] is used to represent the flow of the taken steps as presented in Appendix B. A more detailed description of the selection and exclusion is presented in Section 5.2.3.2 as part of the Data Processing stage of PM^2.

The data quality issues that were found in the HES-APC data set during the data cleaning steps were recorded. **Table 5-3** shows the identified data quality issues with 'L' and 'H' indicators. Each available issue is numbered for quality issue identification and each of them is described in the following section.

Table 5-3 Evaluation of data quality issues for HES-APC data set.

	Missing	**Incorrect**	**Imprecise**	**Irrelevant**
Case				
Event		1. L		2. H
Relationship				
C_attribute			3. H	4. L
Position				
Activity name				
Timestamp	5. H	6. L	7. H	
Resource				
E_attribute	8. L	9. L		

There were nine data quality issues have been found using the HES-APC data set following Bose et al. [154]:

1. Incorrect Event
    a. Incorrect diagnosis by sex (Male): 2,316 episodes from 1,914 patients
    b. Incorrect diagnosis by sex (Female): 236 episodes from 339 patients
2. Irrelevant Event
    a. ICD-10 not related to disease 27,380,144 episodes from 2,277,588 patients.
    b. Repeated diagnostic codes: 34,846,113 codes.

3. Imprecise C_attribute
    a. Patient with multiple date of birth 38,683 patients.
    b. Patient aged less than 18 years old by 1 April 2008 7,466,097 patients.
    c. Unknown/ not specified sex 30,387 patients.
4. Irrelevant C_attribute
    a. Patient with single episode 8,699,
    b. Unfinished spells 325,097 episodes from 309
    c. Only have one discharge 4,912,435 patient.
5. Missing Timestamp
    a. Missing discharge date 16,502,462 episodes from 5,853 patients
6. Incorrect Timestamp
    a. Episode with discharge came earlier than episode start 145 eps from 8 patients
7. Missing E_attribute
    a. Missing epistart 611 episodes from 2 patients
8. Incorrect E_attribute
    a. Episode with incorrect SPELEND 37,873 episodes from 1887
    b. Incorrect episode number 5082 eps from 638 patients.
    c. “Not general“ episode 1139 from 454 patients.

5.1.5 Representativeness

In general, the HES-APC contains records of all patients admitted to NHS hospitals in England including admission data from the private or charitable hospitals paid by the NHS. The HES-APC data set has been made available for research and service evaluation since it allows researchers to follow individual patients over time including its universal coverage. The obtained HES-APC data set for this current work contains admission data from the period of the 2007/2008 financial year until the latest available data in 2016/17. It contains 34 million patients aged 18 years old or above, resulting in 145 million admissions while the estimated total of adult population in England by mid-2016 was 43 million [13]. It means that the number of adult patients in the HES-APC data set covers around 79% of the total adult population of England in mid-2016. Based on the above facts, it is fair to say that the HES-APC data set representative to the adult population of England.

The activity of collecting national data was started since 1987 following a recommendation by the England's Department of Health working group in early 1980 [143]. The main purpose of collecting national hospital activity data is to inform management and planning services.

5.2 Initialisation stages

This first experiment was done to check whether process mining approach can be applied to identify disease trajectories using the HES-APC data set, a larger scale of EHR compared to the data set in Chapter 4's experiment. This current experiment used a random selection of one thousand patients' EHR and the selection process was applied five times to evaluate the robustness of the method and the performance of the discovered disease trajectory models. A 5-fold cross-validation method was used for evaluating the model's performance.

5.2.1 Stage-1: Planning

This experiment addresses the last four project objectives: to check the robustness, representativeness, scalability and to identify disease trajectories of patients with AMI. Four iterations of data processing (Stage-3), mining and analysis (Stage-4), and evaluation (Stage-5) were conducted where each objective was addressed in each iteration. Each iteration used the same source of data set but varied in terms of selection scenarios to answer different research questions. Figure 5.3 illustrate these iterations in the PM^2 framework.

In the first iteration, the process mining method was applied using a random data sample taken from the cleaned population data. The objective of the first iteration is to check the robustness of the method. The same approach was applied in the second iteration where the stratified random sampling was used. The objective of the second iteration was to check the robustness of the process mining approach using a better representation of sampling, and then the method was applied again in the third iteration using the cleaned population data to check the scalability. Following the results from the first three iterations, the process mining method was applied in the fourth iteration to identify disease trajectory using the HES-APC data set where the data was pre-selected by only including patients diagnosed with AMI.

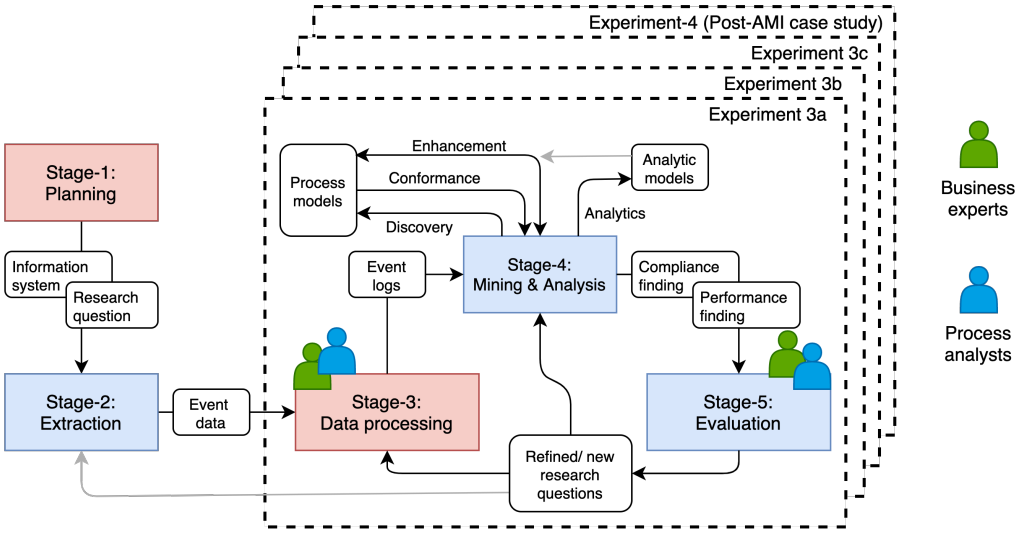


Figure 5.3 PM^2 framework for case study-2.

The scope of analysis for this current work remains similar to the previous experiment in Chapter 4: treating the recorded patients' diagnostic codes as activities to discover the disease trajectory patterns. In this current experiment, primary diagnostic codes were taken from the discharge episodes. A discharge summary was created and submitted by the treating clinician to mark the end of the patient's spell. It contains the principal diagnosis as the main reason for a patient's hospitalisation.

The aim of this current case study was to determine if process mining technique could be implemented to the HES-APC data set – the actual electronic health records, to identify disease trajectories. The aim was composed of a research question where the primary research question of this current case study is *"Can disease trajectories be identified from the HES-APC data set using process mining approach?"* In order to achieve the last four objectives in Chapter 1, four specific research questions were established as follows:

RQ5.1 Is process mining a robust method to identify disease trajectories using random sample data?

RQ5.2 Is process mining a robust method to identify disease trajectories using stratified random sampling data?

RQ5.3 Is process mining scalable in identifying disease trajectories from an actual EHR?

The three research questions above were followed by research questions from the clinical domain:

RQ5.4 What was the progression of diseases after the incidence of AMI?

Following the research questions of RQ5.3 and RQ5.4, frequently-posed questions by medical professionals in process mining were added [50]. The included questions are:

1. What are the most followed paths and how long were the duration of the trajectories?
2. Are there differences in care paths followed by different patient groups?
3. Where are the long waiting time activities in the process?

The team formation in this experiment included experts in epidemiology and a clinician of cardiovascular medicine.

5.2.2 Stage-2: Extraction

5.2.2.1 Extracting event data

Patient records for this current experiment were extracted using Python by applying *pyodbc* – an open-source Python package to establish a connection from Python to SQL Server or any databases that support ODBC connection. Data from the selected columns in Table 5-1 were extracted, containing 145,913,610 rows of data from 34,116,423 patients.

5.2.2.2 Transferring knowledge

Guidelines from the NHS Digital were used as a reference to understand each attribute of the data set and how to conduct data cleaning. These guidelines are the Hospital Episode Statistics Data Dictionary [156] and the HES Autocleanse Dictionary - Accident and Emergency, Admitted Patient Care and Outpatient Care [148]. The selection of the columns was consulted with the experts – the senior epidemiologist and the clinician through meetings and discussions.

As suggested in [156], some of the variables have dependencies on other variables. Information from one variable may be related to several other variables for a better understanding and complete overview of the patient conditions or treatments. Thus, the extraction of primary diagnoses from discharge episodes need to consider other variables. The variables and their respective values are described as follow:

1. Identifiers

   a. Patient identifiers are available as encrypted HES ID resulting in 32 alphanumeric characters. The identifiers are stored in "ENCRYPTED_HESID" variable.
   b. Record identifiers are available as 8 digits numeric and can be up to 14.
2. Episode selection
   a. The episode should have a discharge date i.e., "DISDATE" is not null.
   b. The episode is the end of the spell i.e., the value of "SPELEND" is "Y" (yes).
   c. The episode type is general i.e., the value of "EPITYPE" is "1" (one).
   d. The episode status is "finished", i.e., the value of "EPISTAT" is "3" (three).
   e. The discharge method of the respective episode i.e., "DISMETH" is none of the following values: "not applicable" or "not known".
3. Diagnostic code of primary diagnosis of each episode is recorded under the 'DIAG_01' variable. The international coding standard ICD-10 is used to record the diagnostic codes and standards at various levels of details. In this current experiment, the first three characters of diagnostic codes were used based on the level of the category of the disease.

5.2.3 Stage-3: Data processing

There were three data processing activities conducted in this current work: filtering event data, transforming event log, and event log enrichment. The filtering process includes cleaning and selection activities and is reported using the Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) diagram.

A range of limitations of the data was recognised as described in Section 5.1.3, therefore data preparations were required before the filtering activities were conducted. The preparation activities were:

- adding details to variables containing timestamps as dates were available in the format of months and years ('MMYYYY'). We assumed the dates are the first date of the month which is '1' e.g., a timestamp of '042008' was changed into '01042008' and then converted its data type from a string into *datetime* to allow age calculation using programming language,
- converted data types properly as stated in the data dictionary, and

- calculated the patients' age.

5.2.3.1 Creating views

Using the HES-APC data set, the views were created by including ENCRYPTED_HESID variable as the patient identifier, DIAG_01 variable as the activity name where the primary diagnostic codes are recorded and DISDATE as the time stamp where a patient was discharged from the hospital, and discharge summary was created.

There were multiple variables involved to determine which episode should be considered to construct an event log. In this current work, episodes marked as discharge episodes were taken. The following criteria – based on the NHS HES-APC data dictionary [156], were used to identify if an episode was a discharge episode:

1. The episode was at the end of the spell – i.e. variable SPELEND contains the value of "Y" (yes).
2. The episode type was *general episode*, meaning the episode was not a delivery episode nor a birth episode – i.e. variable EPITYPE contains the value of 1.
3. The episode had a *finished* status, meaning the episode had finished before the end of the HES financial year – i.e. variable EPISTAT contains the value of 3.
4. The discharge method to the respective episode was not of the following values: "not applicable" or "not known" – i.e. variable DISMETH contains values other than 8 or 9.

Once the event log construction was done, event log sorting was conducted to maintain the sequence of diagnostic codes' occurrences. Multiple-level event log sorting was done by:

1. Time when the episode was started – using the variable of EPISTART, then by
2. Episode order number – using the variable of EPIORDER, then by
3. Time when the episode was ended – using the variable of EPIEND, then by
4. Episode identifier – using the variable of EPIKEY.

5.2.3.2 Filtering event data

Filtering event data was done in 16 steps to get the final selection of the event log. Two activities of cleaning and selecting the data were done in alternate ways due to

the limitation of the computing power. The first four cleaning steps were done based on the patient information since the number is smaller than the number of episodes. The steps taken are described as follows:

1. Exclusion of multiple dates of birth.

   Patients recorded with more than one date of birth were excluded from this current work. This step was required to help in calculating the patients' age. Following the NHS Digital's standard [157], the rule for identifying duplicate records failed to be done due to the missing required columns: HOMEADD. Therefore, patients with more than one record of date of birth were excluded.

2. Exclusion by age.

   The DSA contains a description of patients aged less than 18 on 31st March 2009 to be excluded. The exclusion by age in this step was intended to conform to the annual financial period where each period is started every 1st April until 31st March a year after.

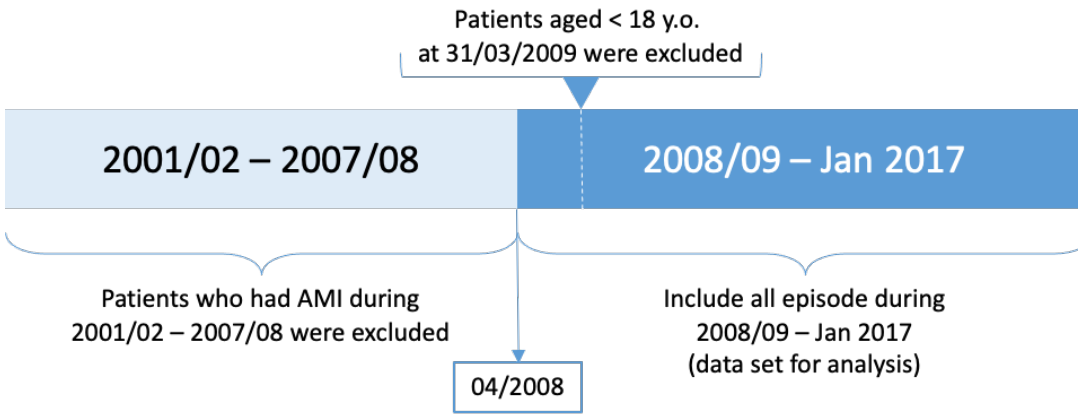


Figure 5.4 Time window selection of the HES-APC

3. Exclusion by sex.

   The column of sex contained four categories. Other categories besides male or female (coded as '1' and '2' respectively) were 'not specified' ('9') and 'not known' ('0'). Patients with the category of '9' or '0' were excluded.

4. Selection by the number of episodes.

   A trajectory is possible if there are two activities involved. This step selected patients with at least two discharge episodes. The filtering activities starting at

this point forward included the discharge episodes data where the number of data was challenging to the computing power. Making sure the utilisation of computing resources to stay under the maximum limit is important. A large number of data were reduced significantly at this current step.

5. Exclusion of missing discharge date.

   As stated in Section 5.2.3.1, this current work used the discharge episodes, therefore the availability of the discharge date was important. Discharge episodes with missing discharge dates were excluded.

6. Exclusion of unfinished spells.

   A discharge episode marks the end of a spell. Each episode has an indicator whether this particular episode is at the end of a spell. If 'yes' then the episode is a discharge episode and the primary diagnostic code in this episode was taken as the activity name.

7. Exclusion of missing episode start date.

   Episode start date is important in facilitating the correct order of the episodes since the time information was only available in months and years. Other information of epiorder, epiend, and epikey were used to determine the order. Epiorder is a sequence number given by the system for each episode within a spell; epiend contains the date of episode ended; and epikey is an identifier created by the HES system for each episode.

8. Exclusion of ICD-10 codes not related to disease.

   The exclusion was done following Jensen et al. (2014) to exclude diagnostic codes that are not categorised as diseases. The categories are based on the ICD-10 chapters of XV and XVI – pregnancy, XVIII – general symptoms and signs not linked to a disease, XIX and XX – external causes, and XXI – administration encounter.

9. Exclusion of incorrect diagnoses by sex.

   An in-depth checking of diagnostic codes was done to make sure that the patients were correctly diagnosed based on their sex. This is to avoid sex-specific diagnostic codes were given to the patients with the opposite sex e.g., a male patient diagnosed with ovarium cancer.

10. Exclusion of non-discharge episodes.

One of the markers to indicate a discharge episode is when the value of the "spelend" attribute is "Y" (yes), which means the episode is the end of a spell. This exclusion is to secure the selection of discharge episodes.

11. Exclusion of discharge date earlier than the episode start date.

    This step was required to maintain the logical sequence of the episode start and discharge episode. A day case episode will have the same episode start date and discharge date, but there were some episodes that have discharge date earlier than the episode start date.

12. Exclusion of invalid episode number

    There are three other values besides the valid sequence number: '98' for *not applicable*, '99' for *not known (a validation error)*, and 'Null' for *not applicable or another maternity event.* Maintaining episodes with valid episode orders is required for maintaining the order of the episode.

13. Selection by episode type.

    Only episodes with the general type were selected. The other possible types are related to delivery, birth, or formally detained under provisions of mental health legislation or long-term psychiatric patients.

14. Excluding non-finished episodes.

    An episode is recorded with a code status of '3' to indicate the finished episode within the HES financial year (before midnight of 31st March) and the patient is discharged. If an unfinished episode is still 'live' at the end of HES financial year, it's recorded with a code status of '1' – *unfinished*, indicating the patient is still receiving treatment in the next financial year or being transferred to another health care provider.

15. Selection of the first occurrence of each diagnostic code.

    This selection step was done to comply with the definition of disease trajectory where only the first occurrences of each diagnostic code were considered. Selecting the first occurrences will avoid a loop in the disease trajectory models.

16. Selection by admission method.

    This selection was made as a result of discussions with the clinical expert where *cataract* was stood out as the most frequent diagnostic codes. According to the clinician's professional judgment, cataract is non-emergency and commonly found in the *elective* admission. The selected admission

methods were *emergency non-elective* except the *labour/ delivery admission* and the *non-emergency transfer* admission method.

17. Repeat step-4.

    Exclusion and selection of episode data result in some patients ending up with a single episode. An iteration of selecting patients with at least two discharge episodes (step-4) was required.

Appendix B shows the filtering report using the STROBE diagram.

5.2.3.3 Filtering 1: exclude rare diagnoses

Among 1,133 unique diagnostic codes, the Pareto approach was used to select diagnostic codes that involved in the 80% of the total episodes. We found 103 unique diagnostic codes that covered 8,825,566 episodes from 3,641,832 patients

The next step was the selection of patients who had at least two diagnostic codes as a minimum requirement to create a sequence of diseases. This selection excluded 733,457 patients with single diagnostic codes resulting 2,908,375 patients with at least two diagnostic codes.

5.2.3.4 Event log transformation

The event log from the previous step were transformed into a pair log, resulting 10,482 unique pairs of diagnostic codes with the lowest frequency of two. The ten most frequent pairs are presented in Table 5-4.

Table 5-4. Ten most frequent pairs of diagnostic codes in HES-APC data set.

Subsequent	Antecedent	Count	%
J44 – Chronic obstructive pulmonary disease (COPD)	J18 – Pneumonia	53,909	1.04
N39 – Disorder of urinary system	J18 – Pneumonia	46,932	0.9
J22 – Acute lower respiratory syndrome	J18 – Pneumonia	37,335	0.72
AMI – Acute myocardial infarction	I25 – Chronic ischaemic heart disease	36,477	0.7
J18 – Pneumonia	J44 – COPD	36,162	0.69
I25 – Chronic ischaemic heart disease	AMI – Acute myocardial infarction	34,341	0.66

J18 – Pneumonia	J22 – Acute lower respiratory syndrome	28,081	0.54
J18 – Pneumonia	N39 – Disorder of urinary system	27,625	0.53
AMI – Acute myocardial infarction	I20 – Angina pectoris	22,356	0.43
…	…	…	…

The result above is obtained after several iterations following discussions with the domain expert. In the previous result, the most dominant pairs of diagnostic codes were Cataract (ICD-10 code H26). The clinical domain expert suspected that the high occurrence of Cataract was coming from the elective admission episodes and suggested to exclude them. Following the suggestions, the admission method criteria in the data dictionary were reviewed. Table 5-5 shows the agreed selected criteria:

Table 5-5 The admission method selection.

ADMIMETH value	Description
	Emergency non-elective admission
21	Accident and emergency
22	GP (direct to hospital provider)
23	Bed bureau (Bed Bureau function provides a single point of access for General Practitioners requiring assessment/admission of emergency patients)
24	Consultant clinic
25	Admission via mental health crisis
28	Other means
2A	A&E
2B	Transfer from emergency
2D	Other emergency admission
	Non-emergency non-elective admission
81	Transfer of any admitted patient from other Hospital

5.2.3.5 Filter-2: selection of strongly associated and significant pairs

Using pair log from the previous step, association measurement was done using Python by implementing the statistic formulas described in Chapter 3. The calculation of *Relative Risk* (RR) was done using 10,482 paired diagnostic codes including the calculation of confidence interval (CI), and the Chi-square test (χ^2) as the significance testing. Table 5-6 shows some parts of the association measurement result.

Table 5-6 The association measurement results of the first 10 paired diagnostic codes.

#	Antecedent (d_1)	Subsequent (d_2)	Count	RR	LCI	p-value
1	J44	J18	53,909	3.07	3.05	0
2	N39	J18	46,932	1.38	1.37	0
3	J22	J18	37,335	1.86	1.84	0
4	AMI	I25	36,477	18.91	18.68	0
5	J18	J44	36,162	3.99	3.96	0
6	I25	AMI	34,341	12.49	12.36	0
7	J18	J22	28,081	2.08	2.06	0
8	J18	N39	27,625	1.21	1.2	2.8E-225
9	AMI	I20	22,356	6.67	6.58	0
10	N39	J22	20,107	1.41	1.39	0
…	…	…		…	…	…

Following the result of association measurement, filtering the paired diagnostic codes was done using the following criteria:

Given the paired diagnostic codes ($d_1 \rightarrow d_2$) pairs with the following criteria were taken:

1. the RR>1 (meaning the incidence of d_2 were likely to happen after d_1) and
2. the CI did not include a value of 1 (lower CI > 1) to avoid the inclusion of pairs with an RR value is equal to 1 (RR=1 meaning the likeliness of the incident of d_2 after d_1 is no difference), and
3. The p-value of χ^2 is less than 0.05

After filtering, the number of paired diagnostic codes was reduced into 2,299 pairs. The selected pairs have a significantly strong association with a p-value less than 0.05. A stricter measurement can be performed for further research such as using a Bonferroni correction where a lower p-value is used for the pair's selection. This

thesis is an exploration of process mining methods to identify disease trajectory using electronic health records, thus using Bonferroni correction is suggested for a deeper investigation.

5.2.3.6 Filter-3: selection by temporal directionality test

Some of the paired diagnostic codes from the previous filtering have the reverse direction ($d_1 \rightarrow d_2$ and $d_2 \rightarrow d_1$). To determine which pair is included in the trajectory, a directionality test is needed.

The 2,299 paired diagnostic codes were measured for temporal directionality using the Binomial test. Pairs with reversed trajectories were tested if one direction is significantly more dominant than the other, but if the test showed non-significant for the pairs, then the pairs were dropped as a trajectory. **Table 5-7** shows the first ten results of the Binomial test.

Table 5-7 The first ten results of Binomial test.

#	Antecedent (d_1)	Subsequent (d_2)	Count	p_value_direction	Significantly greater
1	J44	J18	53,909	0	TRUE
2	N39	J18	46,932	0	TRUE
3	J22	J18	37,335	7.5736E-288	TRUE
4	AMI	I25	36,477	5.14289E-16	TRUE
5	J18	J44	36,162	0	FALSE
6	I25	AMI	34,341	5.14289E-16	FALSE
7	J18	J22	28,081	7.5736E-288	FALSE
8	J18	N39	27,625	0	FALSE
9	AMI	I20	22,356	0	TRUE
10	N39	J22	20,107	5.70809E-53	TRUE

The table above contains pairs of diagnostics that have the inverted directionality (written in red texts). For example, pair number-1 is J44 → J18 and the inverted directionality is J18 → J44 (pair number-5). The selection by directionality test was done by taking the paired diagnostic codes that have a significant value less than 0.05. The column of 'significantly greater' will be set 'True' is the significant value is less than 0.05 or will be set 'False' if the significant value says otherwise. This filtering step resulting 1,055 pairs and the **Table 5-8** below provides the list of ten first paired and temporally directed diagnostics codes including the *RR* and p-values.

Table 5-8 Directional paired diagnostic codes.

#	Antecedent (d_1)	Subsequent (d_2)	Count	RR	p-value RR	p-value Direction
1	J44	J18	53,909	3.07	0	0
2	N39	J18	46,932	1.38	0	0
3	J22	J18	37,335	1.85	0	7.5736E-288
4	AMI	I25	36,477	18.91	0	5.14289E-16
5	AMI	I20	22,356	6.66	0	0
6	N39	J22	20,107	1.41	0	5.70809E-53
7	I50	J18	19,178	1.38	0	4.30293E-12
8	I63	J18	16,484	1.09	4.13E-30	0
9	I48	J18	15,278	1.04	2.51E-06	0
10	I25	I20	15,047	7.69	0	4.5264E-111

The directional paired diagnostic codes were used as the reference to exclude the patients' collection of pairs. Any pairs that do not exist in Table 5-8 were excluded from the patients' collection of paired diagnostic codes. The excluded pairs were kept for further reference.

5.2.3.7 Pair log transformation

The filtered pair log contains 1,173,816 patients with 1,055 pairs of diagnostic codes that are significantly associated with RR>1 and statistically significant in the directionality test. Before using the pair log for further analyses, one more filtering activity was done to make sure that each trace variants are followed by at least two cases. The DISCO was used to conduct the filtering and resulting in a pair log that contains 1,165,885 patients with 8,904 trace variants as the final pair log. This final pair log is referred to as the final cohort and will be used as a comparator in the next analyses.

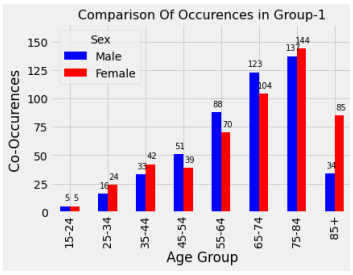
5.2.3.8 Prepared Data set

Two sample data sets were prepared for the current case study that were extracted from the final cohort:

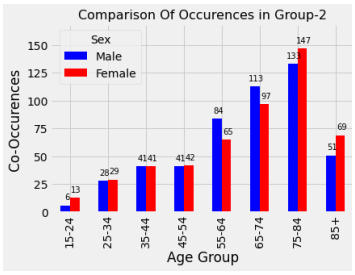
1. Five sets of 1,000 patients were sampled using a simple random selection.
2. Five sets of 1,000 patients were sampled using a stratified random selection.

5.3 Experiment 3a: Process mining of disease trajectory using simple random sampling of patients' EHR

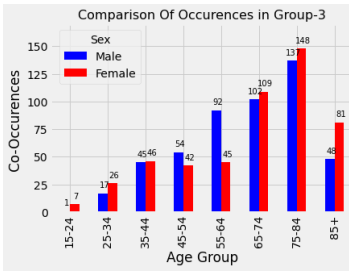
The first experiment for identifying disease trajectories using the HES-APC data set is by implementing the method using randomly selected patients. The random selection with replacement was done five times to create five groups of one thousand patients. The "with replacement" approach in simple random sampling means that every attempt of a randomly selected patient returns the selected patient to the data pool and can be chosen again in the following random attempts. Thus, one patient will have the chance to be selected more than once in 1,000 attempts. This approach will collect 1,000 data but the selected patients are probably less than 1,000. Doing the 1,000 random selection five times will produce five groups of 1,000 data where the number of selected patients in each group is 1,000 patients or less. The results of this selection are presented in Figure 5.5.



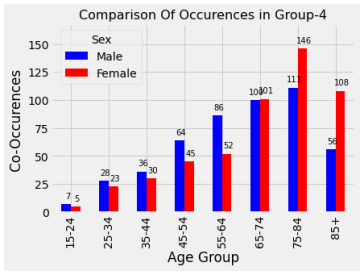
Group-1 (n=1,000)



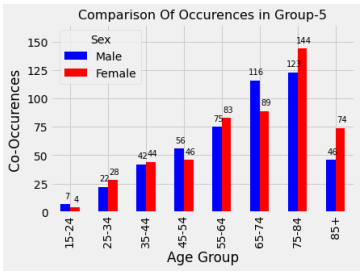
Group-2 (n=1,000)



Group-3 (n=1,000)



Group-4 (n=998)



Group-5 (n=999)

Figure 5.5. Five groups of randomly selection patients.

We took the sampling with replacement because it could preserve the disease trajectory duration by chance. Suppose a sample contains patients with a long duration. In that case, there is more probability that those patients will be included multiple times in the sample, thus preserving the overall duration of the sample.

From the selection results above, the distribution of patients' age, the number of events, and trajectory duration are presented and compared to the final cohort data set containing 1.1 million individuals (see Table 5-9). In this current work, the final cohort data set is used as the "comparator". This is useful for checking the median and IQR of age, number of events, and duration are relatively similar to the comparator.

Table 5-9. The descriptive statistics of the five randomly selected patient groups.

Group-#	N	Age	# Of event	Duration*
		Median (IQR)	Median (IQR)	Median (IQR)
1	1,000	69.5 (55 – 78.25)	4 (4 – 4)	9 (2 – 27)
2	1,000	70 (56 – 80)	4 (4 – 4)	9 (2 – 27)
3	1,000	70 (55 – 79)	4 (4 – 4)	9 (2 – 27)
4	998	70 (55 – 79)	4 (4 – 4)	9 (2 – 27)
5	999	71 (57.5 – 79)	4 (4 – 4)	9 (2 – 27)
Comparator**	1,165,885	71 (56 – 80)	4 (4 – 4)	9 (2 – 27)

* *In months*

** *Final cohort data set*

Figure 5.5 shows the frequency of selected patients in each age category for every group of random data set. These frequencies are relatively like the comparator data set. Similar to the frequency, values in Table 5-9 shows the characteristics of five random data sets relative to the comparator data set. The median and the IQR of age, number of events and duration of each random data set also similar relative to the comparator data set.

The proportion of male and female patient in some of the age group in each random selection group are inconsistent to the comparator data set. The number of female patients in the age group 25-34 years old in group-1 is less than the number of male patients. Meanwhile, the number of female patients in the comparator data set is higher than the male patients. This inconsistency is judged only by looking at the frequency of male and female patients in each age group, let alone the actual proportion of males and females in each age group.

Despite the different proportions between the randomly selected data sets and the final cohort, the frequency of male and female patients in each age group still produces a relatively similar proportion as the final cohort data set. It is fair to say that the random selection produced a relatively representative to the final cohort data set.

Conducting multiple random sampling with replacement to get an exact 1,000 patients in a sample is against the principles of random sampling. The purpose of random sampling with replacement is to get sample data where each member of the population has an equal chance of being selected in the sample. To get the exact number of members in multiple random sampling is better done using the stratified random sampling, which is demonstrated in subchapter 5.4.

5.3.1 Stage-4: Mining and analysis

5.3.1.1 Discovery

The discovery of disease trajectories was done using the five random data sets from the previous stage. Two approaches were used for the discovery stage since each approach has its own benefit in helping to answer the research questions. The first approach was using DISCO – a process mining tool that could provide a range of information including the trace variants and the duration of a disease trajectory which can be exported for further analyses to get the median duration of each trajectory. Each group of the randomly selected patient was loaded into DISCO and then mined for disease trajectory model, trace variants, and durations. The second approach was a discovery using the ProM framework that provides a range of plugins for conformance analysis.

Results from one group out of five groups of random data set are presented in this section as an example.

The data set in Group-1 consist of 101 diagnostic codes and the top 20 is presented in Figure 5.6. The five most frequent diagnostic code is *J18-Pneumonia* and then followed by *N39-Disorder of urinary system, AMI-Acute Myocardial Infarction, J22-Unspecified acute lower respiratory infection,* and *I50-Heart failure*.

The process mining results are presented in Figure 5.7, followed by Table 5-10**Error! Reference source not found.** showing the five most common trajectories, including their respective duration presented in median (IQR).

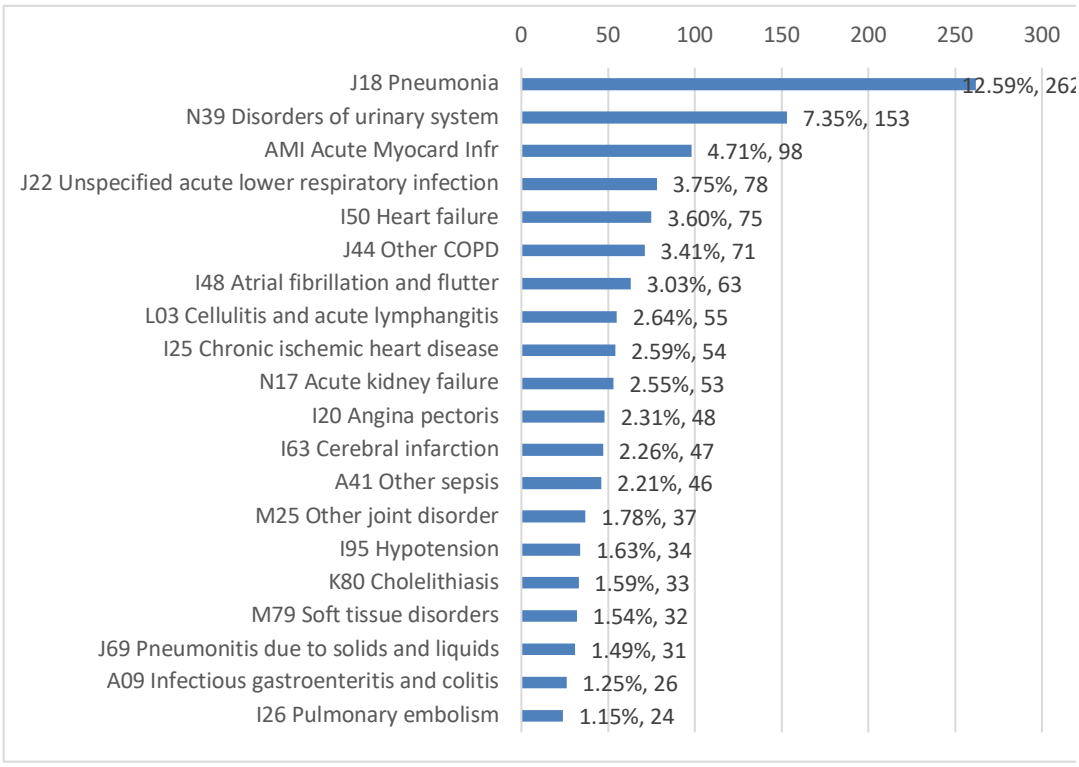


Figure 5.6 The 20 most frequent diagnostic codes in Group-1 data set.

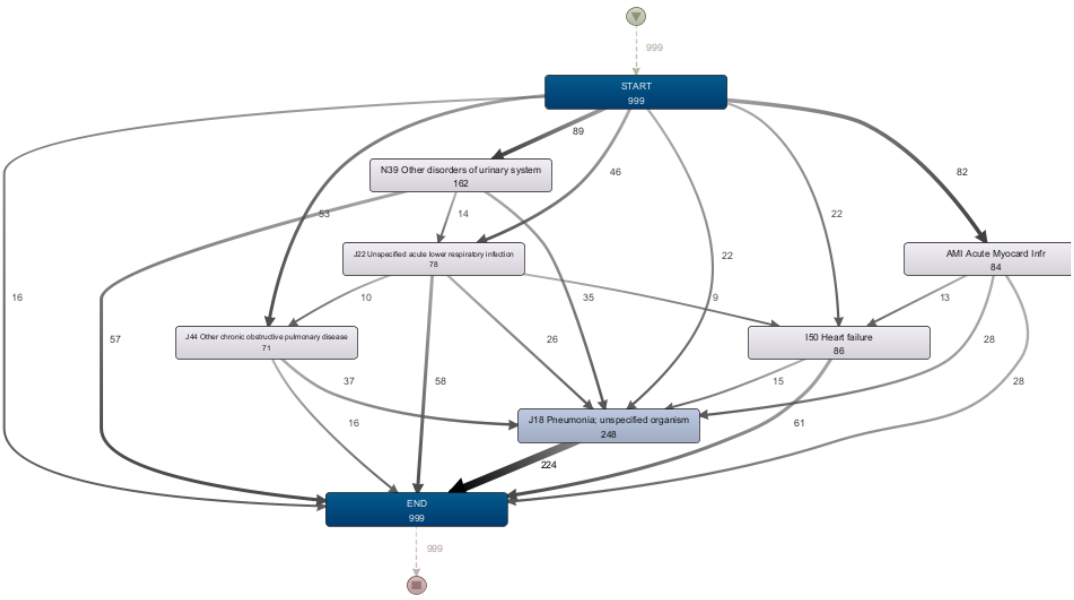


Figure 5.7. Disease trajectory model of the first group of randomly selected patients (n=1,000). The model was generated using DISCO (activities: 5%, paths: 100%).

Table 5-10 Five most common trajectories out of 432 variants from Group-1 of randomly selected patients (n=1,000).

#	Trajectory Variant	N (%)	Median duration (IQR)*
1	START→J44→J18→END	39 (3.9%)	8 (2-20)
2	START→N39→J18→END	34 (3.4%)	10 (1-23)
3	START→AMI→I25→END	33 (3.3%)	0 (0-1)
4	START→J22→J18→END	29 (2.9%)	4 (2-13)
5	START→AMI→I20→END	17 (1.7%)	6 (1-13)

* *In months*

The disease trajectory model presented in Figure 5.7 was generated from the Group-1 data set containing 1,000 randomly selected patients. From the discovered model, the five most common trajectories are presented in Table 5-10 where the most frequent trajectory is *J44-Other chronic obstructive pulmonary disease → J18-Pneumonia.* Another important discovery was the duration of the trajectories. The information of duration for each trajectory is important to show how long the disease has been progressed. Using the median duration and its respective IQR, the shortest and longest trajectories of Group-1 were identified. A trajectory with the shortest median duration was *AMI Acute myocardial infarction → I25 Ischemic heart disease* where the median duration was 0 months (IQR 0-1). The I25 is likely occurred after AMI in less than a month. A trajectory with the longest median duration was *AMI Acute myocardial infarction → I48 Atrial fibrillation* where the median duration was 31 months (IQR 2-43). This longest median duration trajectory was chosen to be reported in this thesis as it has a frequency of five at minimum. This decision was made during one of the meeting sessions with the domain experts as an act of data protection. They suggested not to disclose any data on single patient or any other information if one of the trajectories was only relevant to fewer than five patients, as this was considered potentially identifiable data.

The discovery of disease trajectory models was repeated and applied to the other groups of randomly selected patients. In each group, the most frequent diagnostic codes were identified and compared. The most common diagnostic codes from the combination of five random data sets were identified and then used as a reference in determining the rank of diagnostic codes of each data set. The rank of each diagnostic code in every random group was visualised as in Figure **5.8**.

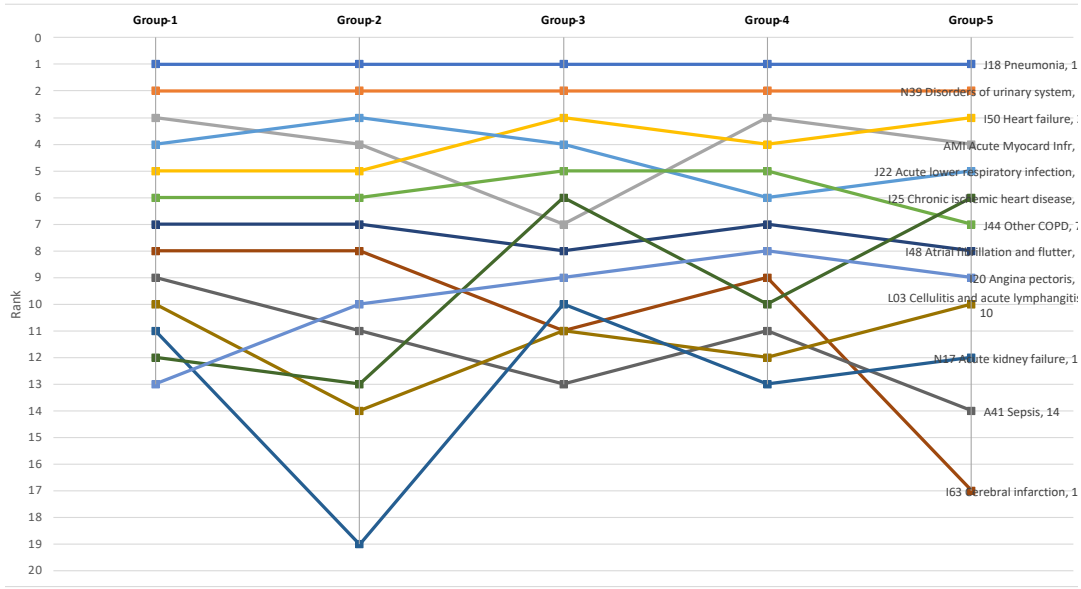


Figure 5.8 The rank of the frequent diagnostic codes for each group.

The calculation of median duration and inter-quartile range (IQR) for each trajectory was done in Python. Both process mining tools of DISCO and ProM, at the time of this thesis being written, are still missing the feature of calculating median duration and the IQR.

The mining activity in this current experiment was repeated using the other four groups of randomly selected patients. The result shows that the patterns of the first four common trajectories in Group-1 are also captured in the rest of the groups (Group-2 to 5) even though the ranking in each group were varied. The pattern of the most common trajectories in each group is presented in Table 5-11.

Table 5-11 The rank mapping of the frequent trajectories from a Simple Random Sampling data sets.

	The rank of traces by frequency				
Trace	**Group-1**	**Group-2**	**Group-3**	**Group-4**	**Group-5**
J44 Other COPD → J18 Pneumonia	1	1	1	1	1
N39 Disorder of urinary system → J18 Pneumonia	2	2	2	3	2
AMI Acute myocardial infarction →I25 Chronic ischaemic heart disease	4	4	5	2	3
J22 Acute lower respiratory syndrome →J18 Pneumonia	3	3	4	5	4
AMI Acute myocardial infarction → I20 Angina pectoris	5	6-10	6-10	6-10	6-10

The patterns of disease trajectories based on the duration are also showing a similar characteristic. The trajectory of *AMI-Acute myocardial infarction* followed by *I25-Chronic ischaemic heart disease* is the shortest duration in each group.

5.3.1.2 Conformance check

There were three out of four measures done for the conformance checking: replay fitness, precision, and generalisation. One measure – the simplicity, was not conducted. Unlike the organisation's process businesses, the nature of the study is not aiming at a simple trajectory. The journey of people getting diseases is personal and cannot be standardised as the cause of a disease is too complex for identification [16]. The conformance check was conducted after the discovery of disease trajectory models using the iDHM plugin in ProM. A single parameter setting was applied to all group of random data set to discover disease trajectories and produce the disease trajectory models. The PetriNet visualisation was chosen as the trajectory models since this type of visualisation is workable to the conformance checking's plugins.
The produced models were then used against their respective event logs to perform conformance checking. Each conformance measure produces a score where the value lies between zero and one (0-1); zero means the quality is poor and one means otherwise. The results of conformance checking from each random group are presented in Table 5-12.

Table 5-12. The conformance checking scores of five randomly selected patient groups.

Group-#	N	Fitness	Generalisation	Precision
1	1,000	0.8702	0.9850	0.7303
2	1,000	0.8871	0.97897	0.8161
3	1,000	0.8983	0.9802	0.7909
4	998	0.8798	0.9794	0.7745
5	999	0.8689	0.9874	0.7300

The overall score of fitness, precision and generalisation is above 0.7, suggesting that the discovered models of each random group are representative to the traces in the event log, generalisable, and precise.

5.3.2 Stage-5: Evaluation

The current experiment investigated the process mining approach to identify disease trajectory models using 1,000 randomly selected patients in the HES-APC data set. The process mining approach was able to produce a disease trajectory model that gave a good result of conformance checking. The method was repeated using four sets of

1,000 randomly selected data and still produced a relatively similar result as the first attempt.

The discovery process on five random data set identified a similar set of four most common trajectories. A relatively similar set of trajectories of diseases with the shortest duration have also been able to be identified.

Further validation was performed using 5-folds cross-validation. A "fold" is equal to one validation data that was taken from the five data sets where the other four data sets become the training data set to build a disease trajectory model. The cross-validation was done five times to make each data set become the validation data. The conformance values of each fold were averaged to get the final result. The results of this current validation are presented in Table 5-13.

Table 5-13 The result of 5-folds cross-validation in Experiment-3a.

#	Training folds	Validation fold	Fitness	Precision	Generalisation
1	2, 3, 4, 5	1	0.9096	0.7829	0.9811
2	1, 3, 4, 5	2	0.9224	0.7245	0.9779
3	1, 2, 4, 5	3	0.9247	0.7637	0.9758
4	1, 2, 3, 5	4	0.9056	0.7186	0.9758
5	1, 2, 3, 4	5	0.9175	0.7575	0.9758
Average			**0.9160**	**0.7494**	**0.9779**

Table 5-13 shows that the average values of fitness, precision, and generalisation from the five folds were high. These results suggest that the disease trajectory models are representative of the traces in the event log, relatively precise to represent only behaviour in the event log, and highly generalisable to reproduce the future behaviours of the trajectories. The above result suggests that the process mining method to identify disease trajectories is robust to sampling.

5.4 Experiment 3b: Process mining of disease trajectory using the stratified random sampling of patients' EHR

This current experiment used a stratified random sampling approach for creating event logs. This second experiment's aim is to see the robustness of the process mining approach in identifying disease trajectories using a more precise sampling

representation, the stratified random sampling, compared to the simple random sampling.

The selection result of the simple random sampling method in the previous experiment (Section 5.3) was not precisely representing the population data, which is the process mining analytics cohort, under more specific attributes. Jensen et al. (2014) [3] suggested that the occurrence of disease has a strong correlation with age and gender, therefore including the strata in patient selection by random is needed.

5.4.1 Stage-3: Data processing

The creation of an event log for this current experiment was done by implementing the stratified random sampling to the process mining analytics cohort data set. Two attributes were included in the sampling process: sex and age group. Following the previous experiment in Section 5.3, the stratified random sampling process was done using the scenario of "with replacement". The sampling process was repeated five times to produce five groups of 1,000 patients. For easier reference, a term of *stratified random groups* is used to refer to the five groups of 1,000 randomly stratified patient selection.

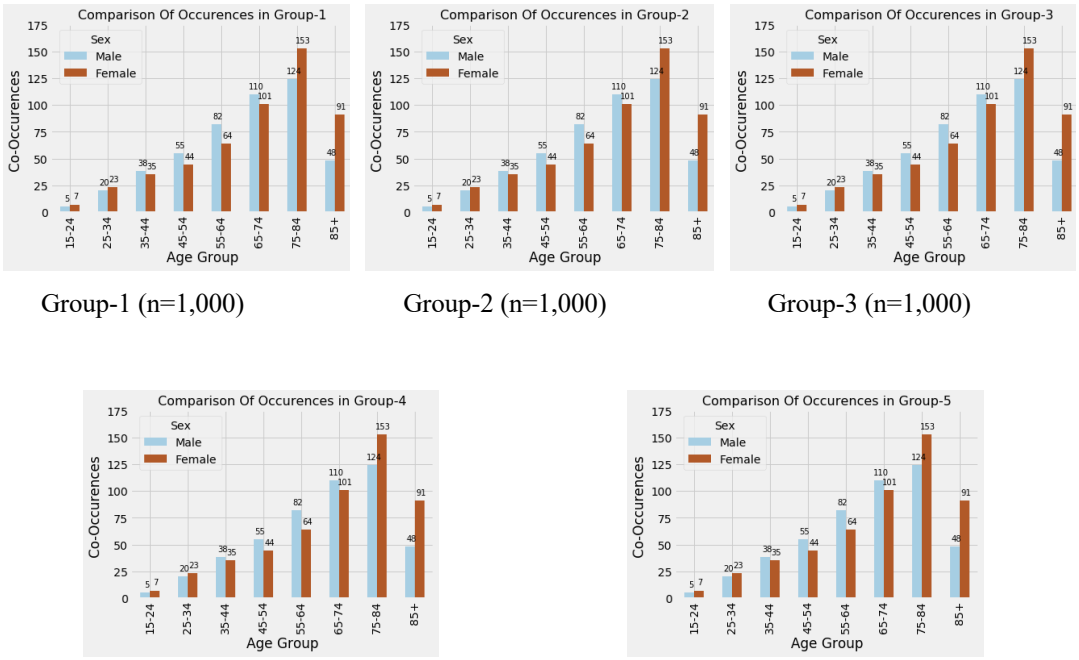
In contrast to the simple random sampling method, the number of sample patients in all stratified random groups is equal to 1,000 patients. The number of samples in every group is consistent. The result of this selection is presented in Figure 5.9 below, while the distribution of age, number of events, and duration for each stratified random group are presented in Table 5-14.

Table 5-14. The descriptive statistics of the five stratified random sampling patient groups.

Group-#	N	Age	# Of event	Duration*
		Median (IQR)	Median (IQR)	Median (IQR)
1	1,000	71 (57 – 80.25)	4 (4 – 4)	9 (2 – 25)
2	1,000	72 (56 – 81)	4 (4 – 4)	9 (2 – 27)
3	1,000	71 (57 – 80)	4 (4 – 4)	9 (2 – 25)
4	1,000	71 (56 – 80)	4 (4 – 4)	9 (2 – 26)
5	1,000	71 (56 – 80)	4 (4 – 4)	8 (2 – 25)
Comparator**	1,165,885	71 (56 – 80)	4 (4 – 4)	9 (2 – 27)

**In months*

***Final cohort data set*



Group-1 (n=1,000) Group-2 (n=1,000) Group-3 (n=1,000)

Group-4 (n=1,000) Group-5 (n=1,000)

Figure 5.9. Five groups of patients selected using stratified random sampling.

5.4.2 Stage-4: Mining and analysis

The activities of discovering disease trajectories in this current experiment were similar to the experiment in Section 5.3. The DISCO was used as the first step in mining disease trajectories, and then the ProM was used for mining disease trajectories and conformance checking.

An example of a disease trajectory model for Group-1 of the randomly stratified selection patient is presented in Figure 5.11. The figure was taken from DISCO and it shows that the most common trajectory is similar to the previous experiment using the simple random sampling: *J44-Other chronic obstructive pulmonary disease* followed by *J18-Pneumonia* covering 31 traces (3.1%) in Group-1, 41 traces (4.1%) in Group-2, 38 traces (3.8%) in Group-3, 45 traces (4.5%) in Group-4, and 44 traces (4.4%) in Group-5. From the discovered model, the five most common trajectories are presented in Table 5-15 where the most frequent trajectory is *J44-Other chronic obstructive pulmonary disease → J18-Pneumonia.*

Table 5-15 Five most common trajectories out of 452 variants from Group-2 of stratified random sampling (n=1,000).

#	Trajectory Variant	N (%)	Median duration (IQR)*
1	START→J44→J18→END	31 (3.1%)	8 (2-20)
2	START→N39→J18→END	31 (3.1%)	8 (2-17)
3	START→AMI→I25→END	28 (2.8%)	0 (0-1)
4	START→J22→J18→END	21 (2.1%)	5 (1-9)
5	START→AMI→I20→END	17 (1.7%)	13 (4-31)

The mining activity was repeated to the other four groups of stratified random sampling data. Five common trajectories occurred in four out of five groups, except in Group-4. Rank is used to sort the trajectories based on the frequency and Table 5-16 shows the variation of the rank in every group.

Table 5-16 The rank mapping of the frequent trajectories from the Stratified Random Sampling data set

	Rank of most frequent sequence				
Trace	**Group-1**	**Group-2**	**Group-3**	**Group-4**	**Group-5**
J44 Other COPD → J18 Pneumonia	1	1	1	1	1
N39 Disorder of urinary system → J18 Pneumonia	3	2	2	2	2
J22 Acute lower respiratory syndrome →J18 Pneumonia	2	3	3	3	4
AMI Acute myocardial infarction →I25 Chronic ischaemic heart disease	4	4	4	4	3
I63 Cerebral infarction → J18 Pneumonia	5	5	5	6-10	5

5.4.3 Simple Random Sampling vs Stratified Random Sampling

The Simple Random Sampling experiment shows that five common trajectories only occur in one group (Group-1). The rest of the group contains four similar common trajectories as in Group-1, but the combinations vary. In contrast, the result of the experiment using Stratified Random Sampling shows that five common trajectories occur in four groups. In comparison, one group contains four common trajectories similar to the rest. The above comparison is presented in Figure 5.10.

	The rank of traces by frequency				
Trace	**Group-1**	**Group-2**	**Group-3**	**Group-4**	**Group-5**
J44 Other COPD → J18 Pneumonia	1	1	1	1	1
N39 Disorder of urinary system → J18 Pneumonia	2	2	2	3	2
AMI Acute myocardial infarction →I25 Chronic ischaemic heart disease	4	4	5	2	3
J22 Acute lower respiratory syndrome →J18 Pneumonia	3	3	4	5	4
AMI Acute myocardial infarction → I20 Angina pectoris	5	6-10	6-10	6-10	6-10

(a) Five common trajectories from simple random sampling experiment.

Trace	Rank of most frequent sequence				
	Group-1	Group-2	Group-3	Group-4	Group-5
J44 Other COPD → J18 Pneumonia	1	1	1	1	1
N39 Disorder of urinary system → J18 Pneumonia	3	2	2	2	2
J22 Acute lower respiratory syndrome →J18 Pneumonia	2	3	3	3	4
AMI Acute myocardial infarction →I25 Chronic ischaemic heart disease	4	4	4	4	3
I63 Cerebral infarction → J18 Pneumonia	5	5	5	6-10	5

(b) Five common trajectories from stratified random sampling experiment.

Figure 5.10 The rank mapping of four most frequent trajectories from Simple Random Sampling and Stratified Random Sampling data sets.

The above result suggests that stratified random sampling produces more stable variation than simple random sampling.

Both experiments produced a similar result regarding the median duration. The trajectory of *AMI-Acute myocardial infarction* followed by *I25-Chronic ischaemic heart disease* (AMI→I25) in both experiments shows the shortest duration (the median duration is less than a month and the IQR is between 0-1 month).

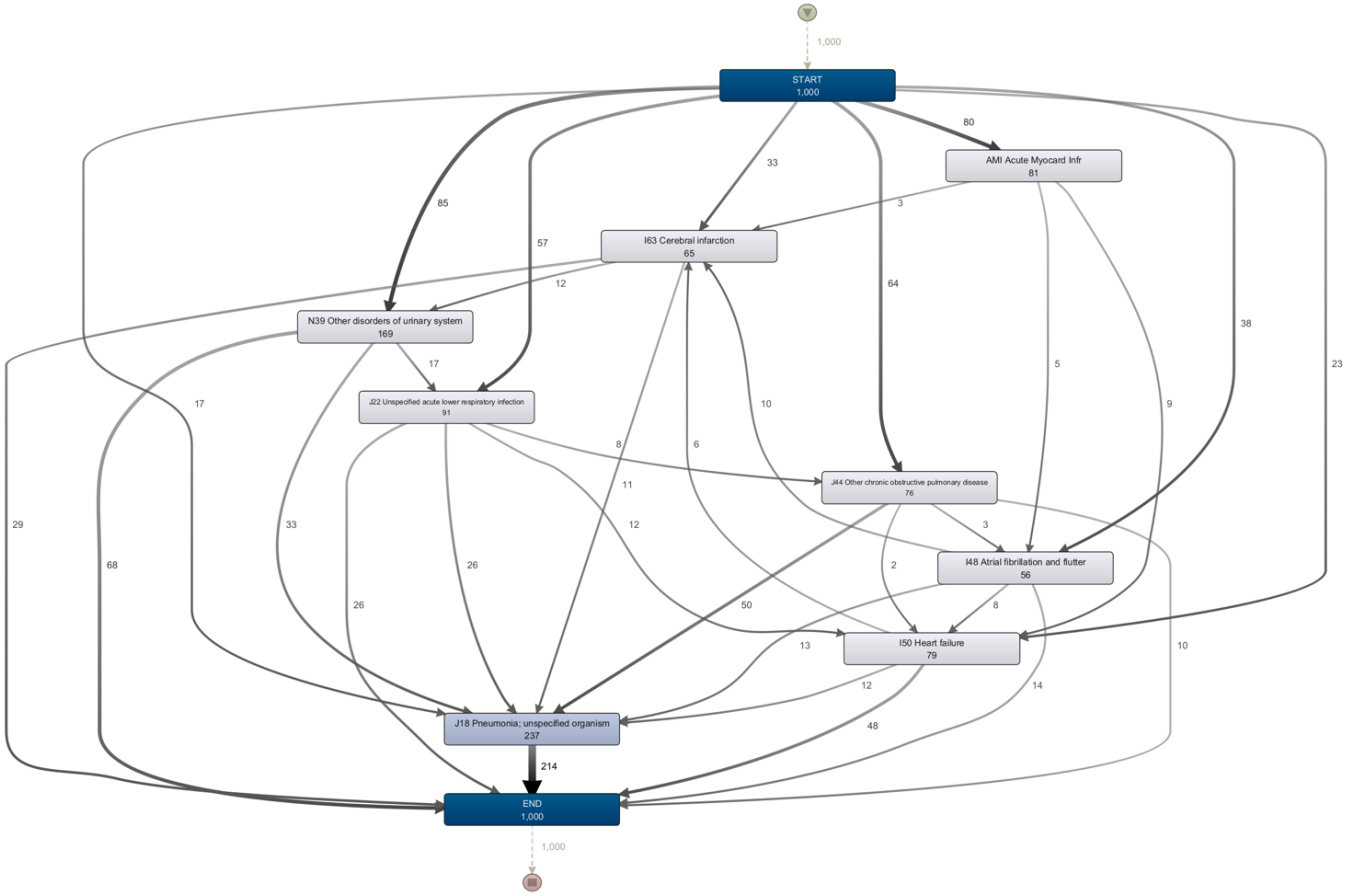


Figure 5.11 An example of disease trajectory model of Group-1 from the stratified random sampling.

5.5 Experiment 3c: Process mining of disease trajectory using the HES-APC data set

Following the results from the previous two experiments, the third experiment was conducted using the 1.1 million cleaned and selected patient data of the process mining analytics cohort. This current experiment is to identify the scalability of the process mining method to identify disease trajectories using an actual EHR.

The size of the data set was scaled up from a sample of 5,000 patients into 1.1 million patients.

5.5.1 Stage-3: Data processing

The data processing activity for experiment 3c has been done to produce the data set for experiment 3a and 3b. The activities were presented in section 5.2.3 and resulting an event log containing 1,165,885 patients with 8,904 trace variants. More details of demographic characteristics are presented in Figure 5.12.

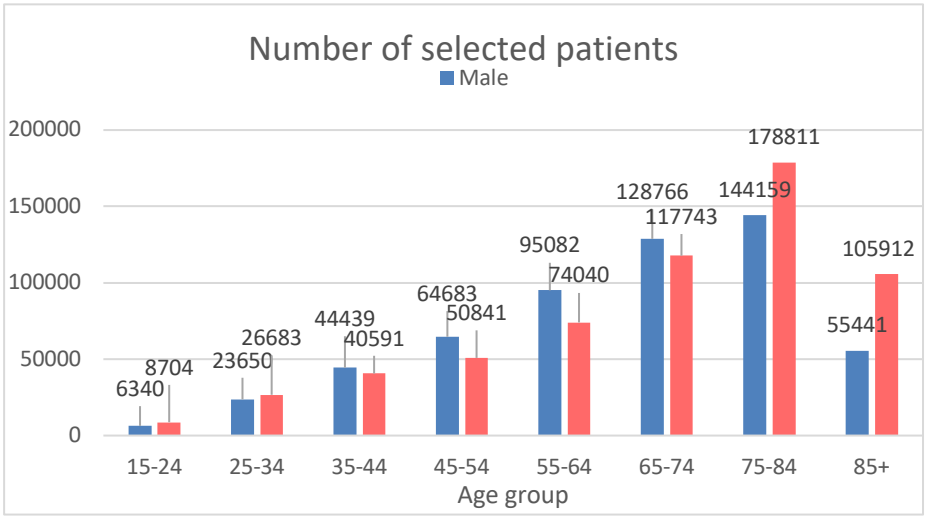


Figure 5.12 Characteristics of the final cohort in Experiment 3c.

5.5.2 Stage-4: Mining and analysis

5.5.2.1 Discovery

The mining and analysis stage resulting the event log was done using two process mining tools: DISCO and ProM. The data description from the DISCO showed that the total number of patients is 1,165,885 with a total of 4,796,642 events resulted from 105 event classes. The median age is 71 years with IQR 56-80 and the median duration is 9 months with IQR 2-27.

Table 5-17 Ten most common trajectories

#	Disease trajectories	N (%)	Median (IQR)*
1	J44 Other COPD → J18 Pneumonia	44,016 (3.78%)	13 (3-30)
2	N39 Disorder of urinary system → J18 Pneumonia	33,176 (2.85%)	9 (2-24)
3	AMI Acute myocardial infarction → I25 Chronic ischaemic heart disease	28,416 (2.44%)	0 (0-0)
4	J22 Acute lower respiratory syndrome → J18 Pneumonia	27,220 (2.33%)	6 (1-21)
5	AMI Acute myocardial infarction → I20 Angina pectoris	17,507 (1.50%)	4 (1-16)
6	I63 Cerebral infarction → J18 Pneumonia	12,905 (1.11%)	13 (4-31)
7	I50 Heart failure → J18 Pneumonia	11,953 (1.03%)	7 (2-21)
8	I48 Atrial fibrillation →J18 Pneumonia	11,908 (1.02%)	13 (3-33)
9	N39 Disorder of urinary system →J22 Acute lower respiratory syndrome	11,329 (0.97%)	10 (3-24)
10	L03 Cellulitis →J18 Pneumonia	11,211 (0.96%)	13 (4,-31)

**In months*

Based on Table 5-17 above, the five most common trajectories in this current experiment are also the common trajectories in the two previous experiments (in Section 5.3 and 5.4). These also include the first five trajectories with the shortest duration.

The mining activity also discovered the shortest and the longest median durations. Trajectory of the shortest median duration with the most case is AMI → I25 (n = 28,416) where the median duration is 0 (zero). The duration of 0 shows that the median duration is less than 30 days due the limited information of the timestamps (only months and years are available). The disease trajectory with the longest median

duration is K35 → K80 → K57 where the median duration is 99 months. **Table 5-18** shows the result of five shortest and longest median duration of disease trajectories.

Table 5-18 The five shortest and longest median duration disease trajectories ordered by median duration in ascending orders.

#	Disease Trajectories	N (%)	Median (IQR)*
1	AMI Acute myocardial infarction → I25 Chronic ischaemic heart disease	28,416 (2.44%)	0 (0-1)
2	AMI Acute myocardial infarction → I24 Other acute ischaemic heart disease	6,793 (0.58%)	0 (0-2)
3	I25 Chronic ischaemic heart disease → I24 Other acute ischaemic heart disease	3,315 (0.28%)	0 (0-1)
4	K92 → K29	2,467 (0.21%)	0 (0-9)
5	K92 → K22	1,584 (0.14%)	0 (0-5)
…	…	…	…
8900	E11 Type 2 diabetes mellitus → E16 Other disorders of pancreatic internal secretion → N39 Other disorder of urinary system → J18 Pneumonia → A41 Other sepsis	3 (~0.0%)	89 (80-93)
8901	I80 Phlebitis and thrombophlebitis → M79 Other soft tissue disorders →I26 Pulmonary embolism → I48 Atrial fibrillation	2 (~0.0%)	89 (85-94)
8902	I80 Phlebitis and thrombophlebitis → M79 Other soft tissue disorders →L03 Cellulitis → N17 Acute kidney failure	3 (~0.0%)	92 (86-95)
8903	F10 Alcohol related disorder → K26 Duodenal ulcer → K22 Other disease of esophagus	3 (~0.0%)	96 (62-96)
8904	K35 Acute appendicitis → K80 Cholelithiasis → K57 Diverticular disease of intestine	3 (~0.0%)	99 (75-100)

When the data set were stratified by sex, the first five most common trajectories of male and female cohorts are presented in Table 5-19.

Table 5-19 Five most common trajectories by sex.

#	Disease Trajectories	N (%)	Median (IQR)*
	Male cohort		
1	J44 Other COPD → J18 Pneumonia	21,609 (3.84 %)	11 (3-28)
2	AMI Acute myocardial infarction → I25 Chronic ischaemic heart disease	20,963 (3.73 %)	0 (0-1)
3	N39 Disorder of urinary system → J18 Pneumonia	13,812 (2.46 %)	8 (2-22)

4	J22 Acute lower respiratory syndrome → J18 Pneumonia	13,074 (2.32 %)	6 (1-20)
5	AMI Acute myocardial infarction → I20 Angina pectoris	11,186 (1.99 %)	5 (1-17)
	Female cohort		
1	J44 Other COPD → J18 Pneumonia	22,407 (3.72 %)	13 (4-31)
2	N39 Disorder of urinary system → J18 Pneumonia	19,364 (3.22 %)	10 (3-25)
3	J22 Acute lower respiratory syndrome → J18 Pneumonia	14,146 (2.35 %)	7 (1-22)
4	AMI Acute myocardial infarction → I25 Chronic ischaemic heart disease	7,453 (1.24 %)	0 (0-1)
5	N39 Disorder of urinary system → J22 Acute lower respiratory syndrome	7,124 (1.18 %)	11 (3-26)

The trajectory of J44 Other COPD → J18 Pneumonia was found as the most common trajectory in both male and female cohorts. By looking at the patterns, the first four most common disease trajectories were similar but different in terms of the order based on the frequency. The trajectory of AMI Acute myocardial infarction → I25 Chronic ischaemic heart disease has occurred as the second most common pattern in the male cohort which happened to 20,963 (3.73%) patients. In contrast with the female cohort, the same trajectory had happened to 7,453 (1.24%) female patients which occurred as the fourth most common trajectory.

Differences in disease trajectory patterns were also found when the event log was stratified by age group. The first five common trajectories in eight groups of cohorts based on age band are presented in Table 5-20.

Table 5-20 Five most common trajectories by age group

(a) 15-24 (n= 14,496)		
#	**Trajectory**	**n (%)**
1	N20 → N13	343 (2.37)
2	K80 → K85	343 (2.37)
3	N20 → N23	326 (2.25)
4	K85 → K86	220 (1.52)
5	I80 → M79	219 (1.51)

(b) 25-34 (n= 49,381)		
#	**Trajectory**	**n (%)**
1	N20 → N13	958 (1.94)
2	M54 → M51	904 (1.83)
3	N20 → N23	904 (1.83)
4	K85 → K86	894 (1.81)
5	I80 → M79	842 (1.71)

(c) **35-44 (n= 83,650)**		
#	**Trajectory**	**n (%)**
1	AMI → I25	2,197 (2.63)
2	F10 → K70	1,811 (2.16)
3	AMI → I20	1,803 (2.16)
4	K85 → K86	1,348 (1.61)
5	I80 → M79	1,244 (1.49)

(d) **45-54 (n= 113,799)**		
#	**Trajectory**	**n (%)**
1	AMI → I25	5,514 (4.85)
2	AMI → I20	3,628 (3.19)
3	J44 → J18	2,395 (2.10)
4	I25 → I20	2,373 (2.09)
5	J22 → J18	1,677 (1.47)

(e) **55-64 (n= 167,137)**		
#	**Trajectory**	**n (%)**
1	J44 → J18	8,346 (4.99)
2	AMI → I25	7,791 (4.66)
3	AMI → I20	4,029 (2.41)
4	J22 → J18	3,223 (1.93)
5	I25 → I20	2,767 (1.66)

(f) **65-74 (n= 244,366)**		
#	**Trajectory**	**n (%)**
1	J44 → J18	14,679 (6.71)
2	AMI → I25	7,737 (3.17)
3	N39 → J18	5,413 (2.22)
4	J22 → J18	5,302 (2.17)
5	AMI → I20	3,565 (1.46)

(g) **75-84 (n= 321,076)**		
#	**Trajectory**	**n (%)**
1	J44 → J18	13,842 (4.31)
2	N39 → J18	13,460 (4.19)
3	J22 → J18	8,790 (2.74)
4	I63→ J18	5,450 (1.70)
5	I50 → J18	5,089 (1.58)

(h) **85+ (n= 159,691)**		
#	**Trajectory**	**n (%)**
1	N39 → J18	10,734 (6.72)
2	J22 → J18	6,381 (4.00)
3	J44 → J18	4,172 (2.61)
4	I50 → J18	3,620 (2.27)
5	N39 → J22	3,373 (2.11)

Of the 1,165,885 patients, there were 562,560 male patients and 601,759 female patients whom the disease trajectories were shared at least by two patients. The first ten most common disease trajectories by sex are presented in **Table 5-21**.

Table 5-21 Ten most common disease trajectories by age group.

(a) Male (n= 562,560)		
#	**Trajectory**	**n (%)**
1	J44 → J18	21,609 (3.84)
2	AMI → I25	20,963 (3.73)
3	N39 → J18	13,812 (2.46)
4	J22 → J18	13,074 (2.32)
5	AMI → I20	11,186 (1.99)
6	I25 → I20	6,947 (1.23)
7	I63 → J18	6,406 (1.14)
8	AMI → I50	5,891 (1.05)
9	I50 → J18	5,684 (1.01)
10	I48 → J18	5,412 (0.96)

(b) Female (n= 601,759)		
#	**Trajectory**	**n (%)**
1	J44 → J18	22,407 (3.72)
2	N39 → J18	19,364 (3.22)
3	J22 → J18	14,146 (2.35)
4	AMI → I25	7,453 (1.24)
5	N39 → J22	7,124 (1.18)
6	I63 → J18	6,499 (1.08)
7	I48 → J18	6,496 (1.08)
8	AMI → I20	6,321 (1.05)
9	I50 → J18	6,269 (1.04)
10	L03 → J18	6,256 (1.04)

The disease trajectory of J44 COPD → J18 Pneumonia occurred as the most common trajectory in both male and female cohorts with the occurrence of 21,609 (3.84%) and 22,407 (3.72) individuals, respectively. The trajectory of AMI Acute myocardial infarction → I25 Chronic ischaemic heart disease became the second most common trajectory in the male cohort, this was in contrast with the female cohort where the same trajectory occurred as the fourth most common trajectory after the N39 → J18 and J22 → J18 came as the second and third.

The diagnostic code of J18 Pneumonia occurred six times as the subsequent event in the ten most common trajectories of the male cohort. In the female cohort, J18 occurred seven times in the first ten most common trajectories.

5.5.2.2 Conformance checking

The conformance checking was done iteratively to identify the optimal parameters that produce the optimal disease trajectory model by examining the combined score of the fitness, precision, and generalisation based on the frequency. The frequency adjustment determines the number of directly-follows relations of activities that were taken. Low-frequency threshold meaning directly-follows relations with lower frequency will be considered. The parameter adjustments were done in two steps. The adjustment made in the first step was done using larger scales to identify the frequency range where the optimum scores are achieved. The second step was the sensitivity analysis where the adjustments were done using a smaller scale until the optimum scores were obtained.

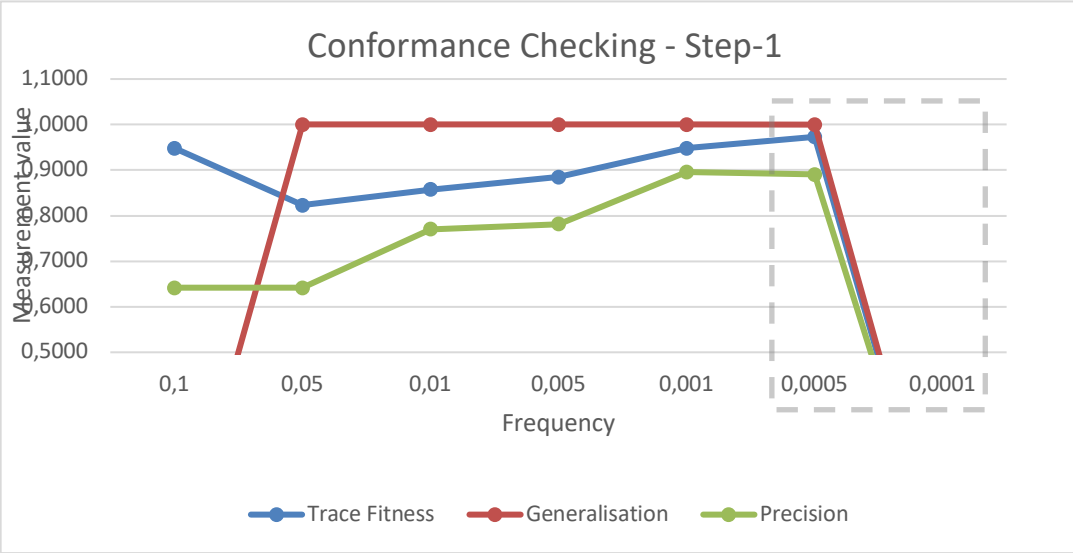


Figure 5.13 Trend of trace fitness, generalisation, and precision in the first sensitivity analysis.

Figure 5.13 shows the relationship between frequency and the measurement values. In the figure, the adjustment on the first sensitivity analysis was done by iteratively reducing the frequency logarithmically started from one tenth combined with a similar pattern that started from five one hundredth, until the ProM raised errors. The lines show the increasing scores of fitness, precision, and generalisation at every point of reduction until the frequency of 0.0001. The changes in the second sensitivity analysis were aimed at the frequency of 0.0005 until 0.0001 and the adjustment was made in a smaller unit of every 0.0001.

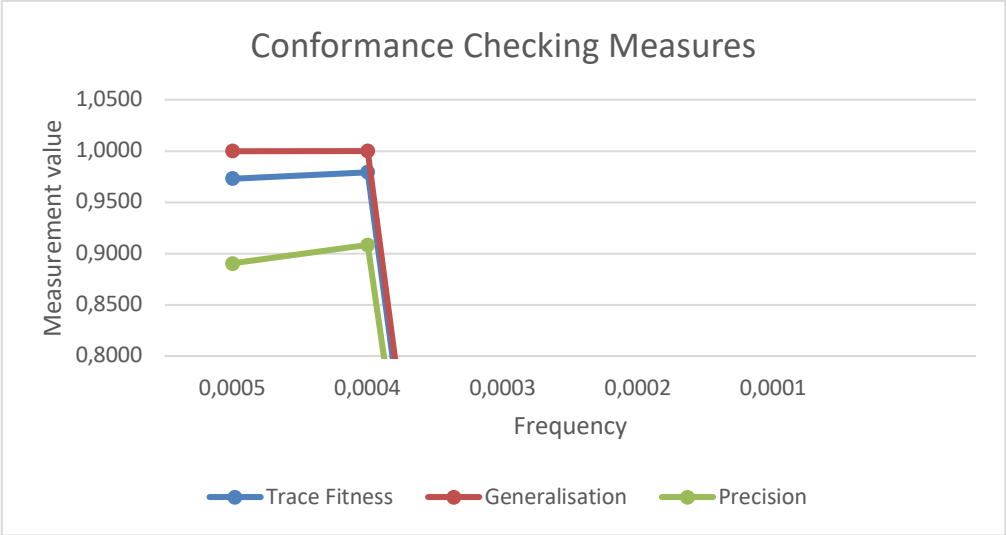


Figure 5.14 Trend of trace fitness, generalisation, and precision in the second sensitivity analysis.

Figure 5.14 shows that the measurement of conformance checking was stopped at the frequency of 0.0004 since any frequency smaller than 0.0004 resulted error in ProM. The ProM tools were unable to produce the Petri net using a frequency lower than 0.0004 and returned an error message "dot.exe has stopped working" as shown in Figure 5.15.

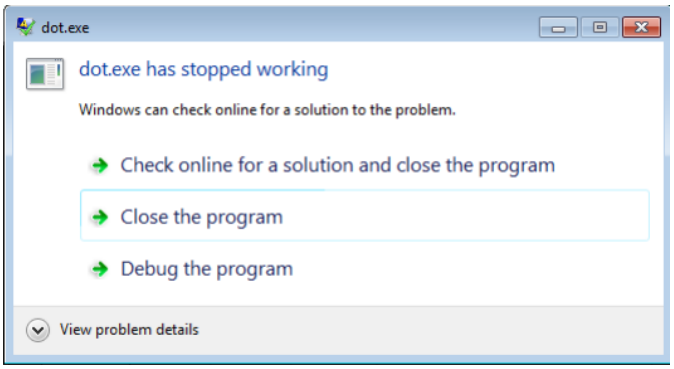


Figure 5.15 The error message when conformance checking measures were done using smaller frequency than 0.0004.

The cause of the error was coming from the visualisation module of ProM where the module had reached its maximum capacity for creating visualisation. Selection using a lower frequency took more data for processing. This investigation is still ongoing but had to be stopped as there was a time limitation for finishing this thesis.

Based on the above limitation, the frequency of 0.0004 was selected as the optimal parameter setting for the current discovery model. The model has a fitness score of 0.9792, generalisation of 0.9999, and a precision of 0.9084, suggesting that the model is representative to the trajectories in the event log.

The obtained findings and results from this current experiment were used to answer the third research question of *RQ5.3 Is process mining scalable in identifying disease trajectories from an actual EHR?* The answer to the question is presented below.

RQ5.3 Is process mining scalable in identifying disease trajectories from an actual EHR?

TThe scalability of the process mining technique in identifying disease trajectories from the HES-APC data set was demonstrated by the results. The common trajectories identified in this experiment were also the common trajectories in the two previous experiments: disease trajectories identification using simple random sampling and using stratified random sampling.

The selection of sample data was repeated five times and yet still produced a relatively similar trajectory. The most common trajectory in this current experiment was occurred as one of the top four most common trajectories in both the experiment using the simple random sampling and using the stratified random sampling.

The median duration and the IQR of this current experiment quantify the results from the two previous experiments:

- The median (IQR) duration of the current experiment was 9 months (2-27), while
- the median (IQR) duration of the five simple random sampling were 9 months (2-27), and
- the median (IQR) duration of the five stratified random sampling was either 8 to 9 months and the third quartile were in a range of 25-27 months.

The median (IQR) of patients' age in this current experiment was also quantify the results from the two previous experiments:

- The median (IQR) patients' age in the current experiment was 71 years old (56-80), while
- The median (IQR) of patients' age of the five simple random samplings was 71 to 72 with the first quartile of IQR varying between 56-57 years old, and the third quartile varying between 80 to 81 years old.

- The median (IQR) of patients' age of the five stratified random sampling was 69.5 to 71 with the first quartile of IQR were varied between 55-57.75 years old, and the third quartile varying between 78.25 to 80 years old.

Despite the limitation on the capacity in handling visualisation, the results from the conformance checking suggested that the process mining technique is scalable for disease trajectories identification.

5.5.3 Stage-5: Evaluation

The evaluation was done through discussions where the clinical expert was consulted about the method and the result. The application of the method for general hospitalisation was measured and an additional evaluation was done using the 5-folds cross-validation. The validation result is presented in **Table 5-22**.

Table 5-22 The result of 5-folds cross-validation.

#	Training folds	Validation fold	Fitness	Generalisation	Precision
1	2, 3, 4, 5	1	0.9447	0.9994	0.8925
2	1, 3, 4, 5	2	0.9468	0.9916	0.8881
3	1, 2, 4, 5	3	0.9462	0.9985	0.8909
4	1, 2, 3, 5	4	0.9476	0.9994	0.8862
5	1, 2, 3, 4	5	0.9483	0.9952	0.8879
Average			**0.9467**	**0.9968**	**0.8891**

Table 5-22 shows that the average score of fitness, generalisation, and precision from the 5-folds cross-validations were high. The high fitness suggests that the disease trajectory model was representative of the trajectories in the event log, the high generalisation means the model still allows more variety of possible future trajectories, and the high precision means the model does not allow sequences of diagnostic codes that is unlike the observed sequences in the event log.

The limitation of analysing the HES-APC data set was the incomplete timestamp data. The duration of disease trajectories is presented using the number of months since the timestamps were available in months and years.

5.5.4 Discussion

Domain experts were involved since the beginning of this study: (1) Chris P. Gale, a professor of Cardiovascular Medicine and Honorary Consultant Cardiologist at the University of Leeds and (2) Dr Marlous Hall, a senior epidemiologist of cardiovascular epidemiology at the University of Leeds. They shared their knowledge and helped in every stage of the study. In the planning stage, their help was useful in understanding the study's context, the EHR, and clinical terminologies and identifying the possible sources for better analysis. During the data processing stage, experts' knowledge and experience helped the study to gain better accuracy regarding the selection criteria. For example, the selection results at the early stage of the study show that the number of diagnostic codes related to eye disease (cataract) is dominant. According to the cardiologist, this disease is common, and the number is increasing. This statement is supported by the previous study by Alrawashdeh et al. (2021), showing that the number of hospitalisation due to eye problems is growing in the last two decades [158].

This current work has helped the clinician understand certain diseases better. One example is the short-duration trajectory of trajectories that begins with acute myocardial infarction. This study found that the shortest median duration is "zero" months (IQR 0-0) (less than a month) for the trajectory of AMI→I25 (acute myocardial infarction followed by ischaemic heart disease). This study also opens opportunities for future works to understand the progression of comorbidities."

5.6 Experiment 4: Process mining of disease trajectory of patients with *acute myocardial infarction*

The experiment in Section 5.5 showed that the process mining technique is workable for identifying disease trajectories using an actual EHR. In this current experiment, the method was implemented to identify the pattern of diseases after the occurrence of acute myocardial infarction (AMI).

5.6.1 Stage-3: Data processing

The selection criteria for this current experiment were imposed to only select patients who had AMI incidents. The sequence of diagnostic codes for disease trajectories began with AMI and was followed by other diagnostic codes that came after the first

AMI. Any recurrence of AMI was dismissed as including the second occurrence or more will not comply with the definition of the disease trajectory. Only the first occurrence of each disease will be considered.

The selection was conducted using a process mining tool DISCO by exploiting the Filter feature. From 1,165,885 patients in the previous experiment (Section 5.5), three filtering steps were done to get the post-AMI cohort:

1. Filtering by follower

   The first filtering was done to select trajectories that include AMI. Other trajectories that do not have AMI were excluded. The filtering was done based on the 'Activity'. The *Reference event value* was set as "START" and the *Follower event value* was set as "AMI". There were 1,069,933 (91.77%) patients excluded and 95,952 (8.23%) patients remained with 398,751 events.

2. Filtering by endpoints

   The second filtering used "Trim the longest" option to produce patients' trajectories that begin with AMI. This filtering activity removed the diagnostic codes that occurred before the AMI event. From 398,751 events, 97,503 events were trimmed resulting in 301,248 events from 95,952 patients.

3. Filtering by follower

   The previous filtering activity produced a trajectory where there was only one AMI event left and then followed by an END event. The sequence is no longer a trajectory. Using the filtering by the follower, the sequences of one AMI and END events were excluded from the data set. The *reference event value* was set to "AMI" and the *follower event value* was set to "END", with the *"reference event must be"* was set into *"never directly followed",* which means only select the trajectories that begin with AMI and followed by any other events instead of END event. This third filtering activity resulting in 94,751 patients with 298,846 events, and 384 trace variants remained.

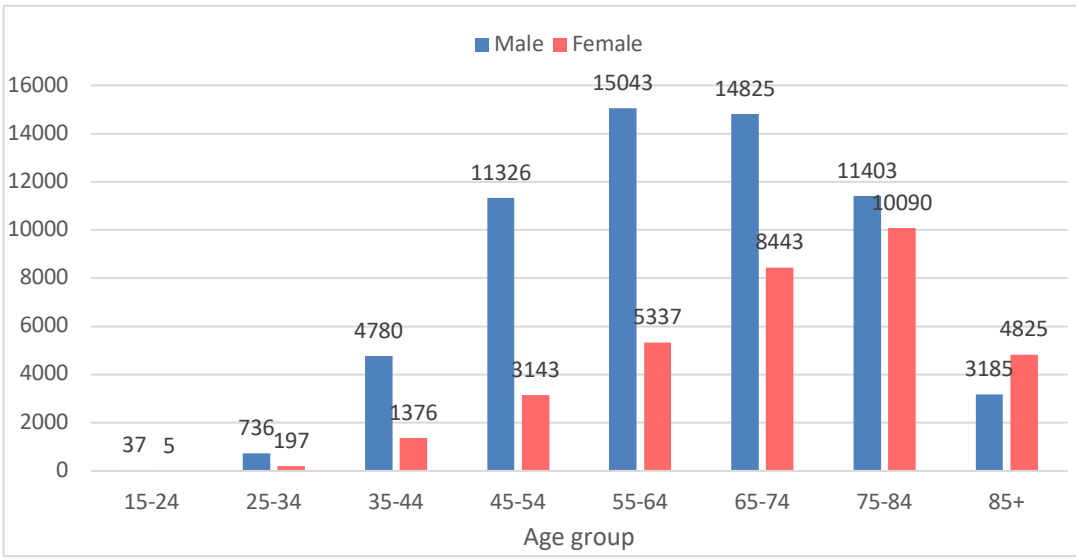
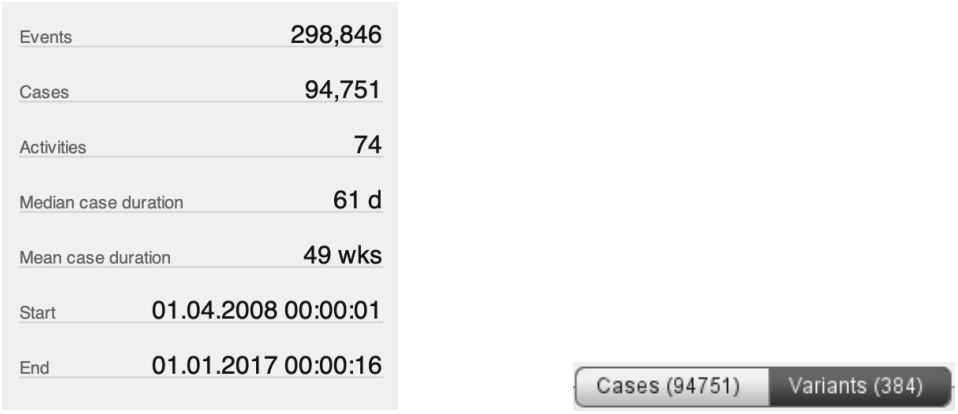


Figure 5.16 Patient distribution based on sex and age group.

5.6.2 Stage-4: Mining and analysis

5.6.2.1 Discovery

Similar to the three previous experiments, the discovery activities were done using two process mining tools: DISCO and ProM. The descriptive result from DISCO showed that there were 298,846 events from 94,751 patients that involved 74 unique activities. In this term, activities are the unique diagnostic codes. The median duration is 61 days (approximately two months) and the mean duration is 49 weeks (approximately 11 months). Figure 5.17 shows the DISCO's statistical result.



(a) (b)

Figure 5.17 The statistical information of (a) the post-AMI event log and (b) the number of case and trace variants in DISCO.

The number of trace variants is 384 where the ten most common post-AMI trajectories are presented in Table 5-23.

Table 5-23 Ten most common post-AMI trajectories.

#	Trajectories	Median Duration (IQR)	N (%)	(%) Cumulative
1	AMI → I25 Ischemic heart disease	0 (0-1)	28,491 (30.07)	30.07
2	AMI → I20 Angina pectoris	4 (1-16)	17,554 (18.53)	48.6
3	AMI → I50 Heart failure	3 (1-13)	10,784 (11.38)	59.98
4	AMI → I24 Other acute ischemic heart disease	0 (0-2)	6,826 (7.20)	67.18
5	AMI → I48 Atrial fibrillation	7 (2-25)	4,563 (4.82)	72
6	AMI → I63 Cerebral infarction	1 (2-32)	4,453 (4.70)	76.7
7	AMI → I95 Hypotension	7 (2-25)	2,051 (2.16)	78.86
8	AMI → I25 Isch. heart disease→ I20 Angina pectoris	9 (2-25)	1,965 (2.07)	80.93
9	AMI → K21 Gastro-oesophageal reflux disease	8 (2-24)	1,926 (2.03)	82.96
10	AMI → I47 Paroxysmal tachycardia	4 (1-19)	1,509 (1.59)	84.55

Table 5-23 shows the ten most frequent trajectories that begin with AMI. The table contains the identification of each variant which available at the "Variant" column. The "Trajectories" column contains the subsequent diagnostic codes after AMI. The "Length" column contains the length of each trajectory. The "Duration" column contains the total duration of each trajectory in months and the median and the IQR (inter quartile range) are presented. The duration is the number of months between AMI and the last diagnostic code in the sequence. The "N (%)" column contains the frequency and the percentage of patients that followed the respective trajectory, and the "(%) Cumulative" represents the percentage of the cumulative frequency of the first ten most common trace variants.

The most frequent diagnostic code after AMI was I25-ischemic heart disease. There were 28,491 (30.07%) patients who had followed the trajectory where the median (IQR) duration was 0 (0-1). The zero month means that the median duration between AMI and I25 was less than a month.

The first eight frequent trajectories covered more than 80% of the cases. The eight diagnostic codes that frequently followed AMI were *I25-Ischemic heart disease*, *I20-*

Angina pectoris, I50-Heart failure, I24-Other ischemic heart disease, I48-Atrial fibrillation, I63-Cerebral infarction, I95-Hypotension, and *I25-ischemic heart disease* followed by *I20-Angina pectoris*.

From the 384 trajectories, there were 99 trajectories where each trajectory was followed by two patients. The 99 trajectories were constructed from the combination of 61 diagnostic codes with the median length of event = 4 (IQR: 4 to 5). The example of five exceptional trajectories is presented in **Table 5-24**.

Table 5-24 Five example of post-AMI exceptional trajectories

Variant index	Exceptional trajectories
286	AMI Acute Myocard Infr→ I48 Atrial fibrillation and flutter→ J90 Pleural effusion; not elsewhere classified→ J22 Unspecified acute lower respiratory infection
287	AMI Acute Myocard Infr→ I48 Atrial fibrillation and flutter→ I95 Hypotension→ M25 Other joint disorder; not elsewhere classified
288	AMI Acute Myocard Infr→ I48 Atrial fibrillation and flutter→ J18 Pneumonia; unspecified organism→ C34 Malignant neoplasm of bronchus and lung
289	AMI Acute Myocard Infr→ I48 Atrial fibrillation and flutter→ I49 Other cardiac arrhythmias→ I50 Heart failure
290	AMI Acute Myocard Infr→ I48 Atrial fibrillation and flutter→ I47 Paroxysmal tachycardia→ I95 Hypotension

From the exceptional trajectories in Table 5-24, the ten most frequent diagnostic codes following AMI are presented in Figure 5.18.

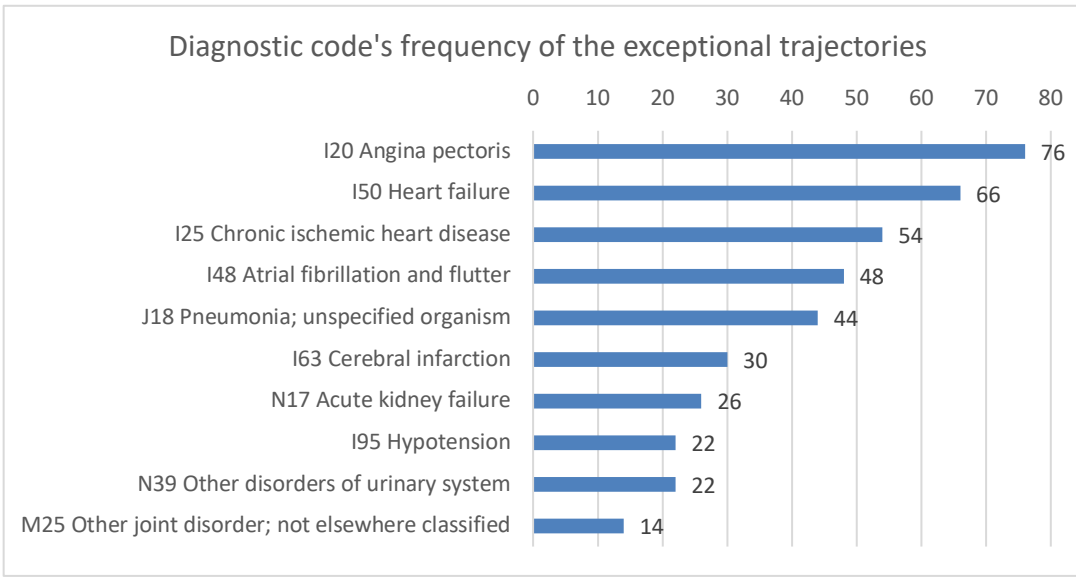


Figure 5.18 Ten most frequent diagnostic codes of the exceptional trajectories

The discovery of disease trajectories was also being identified by different groups of patients. The categories were sex, mortality status, and age group. The five most common trajectories followed by both male and female patients were similar including the order of the rank based on the frequency. These trajectories are *AMI→I25 Ischemic heart disease, AMI→I20 Angina pectoris, AMI→I50 Heart failure, AMI→I24 Other ischemic heart disease,* and *AMI→I48 Atrial fibrillation*.

A similar set of trajectories between male and female patients have also happened in the five shortest durations, but the order of the 4th and 5th place of male patient group was flipped in the female patient group. Table 5-28 and Table 5-29 show the the five most common trajectories and the five shortest durations of trajectories by sex, respectively.

Disease trajectories were also identified for different groups of age. The five most common disease trajectories were identified from each group except in the group of 15-24 years old since there were only four trajectories that shared by at least two patients. The following tables are examples of the five most common trajectories by age group 55-64 (Table 5-25) and 75-84 years old (Table 5-26).

Table 5-25 The example of five most common trajectories in group 55-64 years old (n= 20,295)

Trajectory	Median age (IQR) = 60 (57-62)			
	N	%	Cumulative %	Median duration (IQR)*
AMI→I25 Chronic isch. heart disease	7,803	38.45	38.45	0 (0-1)

AMI→I20 Angina pectoris	4,033	19.87	58.32	0 (0-1)
AMI→I24 Other isch. heart disease	1,619	7.98	66.30	0 (0-4)
AMI→I50 Heart failure	1,209	5.96	72.25	1 (0-9)
AMI→I48 Atrial fibrillation and flutter	923	4.55	76.80	2 (1-3)

The first three trajectories from Table 5-25 are AMI→I25 Chronic ischemic heart disease, AMI→I20 Angina pectoris, and AMI→I24 Other ischemic heart disease occurred as the five most common trajectories in under 75 years old age group. In the age group of 75 years old and above, the trajectory of AMI→I24 Other ischemic heart disease was no longer in the top five but was replaced by AMI→I63 cerebral infarction. The trajectories in Table 5-26 were the five most common trajectories in the last two age groups: 75-84 and 85+ years old.

Table 5-26 The example of five most common trajectories in group 75-84 years old (n= 21,402)

Trajectory	Median age (IQR) = 77 (75-79)			
	N	%	Cumulative %	Median duration (IQR)*
AMI→I25 Chronic isch. heart disease	4,299	20.09	20.09	0 (0-1)
AMI→I50 Heart failure	4,066	19.00	39.09	3 (1-14)
AMI→I20 Angina pectoris	3,063	14.31	53.40	4 (1-14)
AMI→I63 Cerebral infarction	1,417	6.62	60.02	12 (3-29)
AMI→I48 Atrial fibrillation and flutter	1,282	5.99	66.01	7 (2-25)

The median duration of *AMI→I25 Chronic ischemic heart disease* and *AMI→I24 Other ischemic heart disease* were both zero months with an IQR 0 to 1 month in all age groups. Both I25 and I24 were occurred less than a month since the first incident of AMI and this had happened to patients in all age group. These identified trajectories are conforming the finding of AMI as a common early indication of ischemic heart disease [159].

Of the 94,751 patients there were 26,350 (27.8%) deaths during the time window selection, April 2008 until January 2017. Of those 26,350 deaths, 14,630 (55.52%) are male patients, and 11,720 (44.48%) are female patients. The median duration of the 26,350 deaths was 3 months (IQR: 1-15) while the median duration of the censored patients was 1 month (IQR: 0-13). Table 5-27 shows the five most common trajectories for each mortality status that shared by at least two patients.

Table 5-27 The example of five most common trajectories of patients based on mortality status.

Trajectory	Median age (IQR) = 79 (71-84)			
	N	%	Cumulative %	Median duration (IQR)*
A. Group of patients with death outcome (n=26,266)				
AMI→I50 Heart failure	6,647	25.31	25.31	3 (1-11)
AMI→I25 Chronic isch. heart disease	4,378	16.67	41.97	0 (0-1)
AMI→I20 Angina pectoris	3,233	12.31	54.28	3 (0-10)
AMI→I63 Cerebral infarction	2,037	7.76	62.04	9 (2-25)
AMI→I24 Other isch. heart disease	1,407	5.36	67.40	0 (0-1)
B. Group of survived patients at censor date (n= 68,316)				
AMI→I25 Chronic isch. heart disease	24113	35.30	35.30	0 (0-1)
AMI→I20 Angina pectoris	5419	7.93	43.23	0 (0-2)
AMI→I24 Other isch. heart disease	409	0.60	43.83	1 (0-9)
AMI→I50 Heart failure	555	0.81	44.64	1 (0-9)
AMI→I48 Atrial fibrillation and flutter	183	0.27	44.91	1 (0-3)

**In months*

Table 5-28 Five most common post-AMI trajectories by sex.

Trajectory	Male (N = 61,273)					Female (N = 33,340)				
	Rank	N	%	Cumulative %	Median duration(IQR)*	Rank	N	%	Cumulative %	Median duration (IQR)*
AMI→I25 Chronic isch.heart disease	1	21,021	34.27	34.27	0 (0-1)	1	7,470	22.35	22.35	0 (0-1)
AMI→I20 Angina pectoris	2	11,213	18.28	52.55	5 (1-17)	2	6,341	18.98	41.33	4 (1-14)
AMI→I50 Heart failure	3	5,911	9.64	62.19	3 (1-14)	3	4,873	14.58	55.91	3 (1-12)
AMI→I24 Other isch.heart disease	4	4,550	7.42	69.61	0 (0-2)	4	2,276	6.81	62.72	0 (0-2)
AMI→I48 Atrial fibrillation	5	2,574	4.2	73.81	7 (2-25)	5	1,989	5.95	68.67	7 (2-24)

Table 5-29 First five shortest post-AMI trajectories by sex.

Trajectory	Male (N = 61,273)					Female (N = 33,340)				
	Rank	N	%	Cumulative %	Median duration (IQR)*	Rank	N	%	Cumulative %	Median duration (IQR)*
AMI→I25 Chronic isch.heart disease	1	21021	34.31	34.31	0 (0-1)	1	7470	22.41	22.41	0 (0-1)
AMI→I24 Other isch.heart disease	2	4550	7.43	41.73	0 (0-2)	2	2276	6.83	29.23	0 (0-2)
AMI→I25 Chronic isch.heart disease →I24 Other isch.heart disease	3	505	0.82	42.56	1 (0-9)	3	187	0.56	29.79	1 (0-4)
AMI→I35 Nonrheumatic aortic valve disorders	4	486	0.79	43.35	1 (0-9)	5	330	0.99	30.78	1 (0-2)
AMI→I25 Nonrheumatic aortic valve disorders →J90 Pleural effusion	5	181	0.00	43.65	2 (1-4)	4	44	0.13	30.91	2 (0-11)

The first five shortest disease trajectories based on median duration in the post-AMI event were presented in **Table 5-30**, while the longest median duration in **Table 5-31**.

Table 5-30. Five shortest disease trajectories based on median duration.

Variant index	Trajectories	N (%)	Median (IQR) (months)	Cumulative (%)
1	AMI→I25 Chronic isch. heart disease	28,491 (30.07)	0 (0-1)	30.07
4	AMI→I24 Other acute isch.heart disease	6,826 (7.20)	0 (0-2)	37.27
16	AMI→I25 Chronic isch. heart disease →I24 Other acute isch.heart disease	692 (0.73)	1 (0-6)	38.00
14	AMI→I35 Nonrheumatic aortic valve disorders	816 (0.86)	2 (0-10)	38.87
26	AMI→I25 Isch. heart disease →J90 Pleural effusion	225 (0.24)	2 (1-4)	39.10

Table 5-31 First five longest post-AMI disease trajectories based on median duration.

Variant index	Trajectories	N (%)	Median (IQR)*
305	AMI→I20 Angina pectoris →G45 Transient cerebral isch.attacks & Related syndromes →N39 Other disorders of urinary system →J18 Pneumonia	2 ($2.11x10^{-5}$)	86.5 (79.75-93.25)
361	AMI→D50 Iron deficiency anemia →I50 Heart failure →N17 Acute kidney failure	2 ($2.11x10^{-5}$)	80.5 (75.25-85.75)
257	AMI→I25 Chronic isch. heart disease →I44 Atrioventricular and left bundle-branch block →J18 Pneumonia	3 ($3.16x10^{-5}$)	80 (63.5-82)
325	AMI→I20 Angina pectoris →K40 Inguinal hernia →K59 Other functional intestinal disorders	2 ($2.11x10^{-5}$)	78.5 (73.25-83.75)
160	AMI→I95 Hypotension →N39 Other disorders of urinary system →J22 Unspecified acute lower respiratory infection	7 ($7.38x10^{-5}$)	77 (24.5-87.5)

5.6.2.2 Conformance checking

The similar approach as described in Section 5.5.2.2 was performed in this current experiment to check the conformance of the discovered disease trajectory model. The conformance checking was done in an iterative approach to obtain the optimal parameters for discovering the disease trajectory model. The iteration was performed to iterate the adjustment made on the percentage of the directly-follows of the iDHM process discovery. The lower the percentage means more directly-follows were involved and more properties were also included to increase the complexity of the model.

In the first iteration, adjustments were made using the combination of logarithmic scale starting from 0.1 and half of the logarithmic number. For example, it begun with 0.1 and then followed by the half of 0.1 which is 0.05 and then 0.01 followed by 0.005, and so on until 0.00005 as the limit before the system raised an error. The conformance checking results using these scales are presented in Figure 5.19 below with a modification on the lower limit of the y-axis for better presentation.

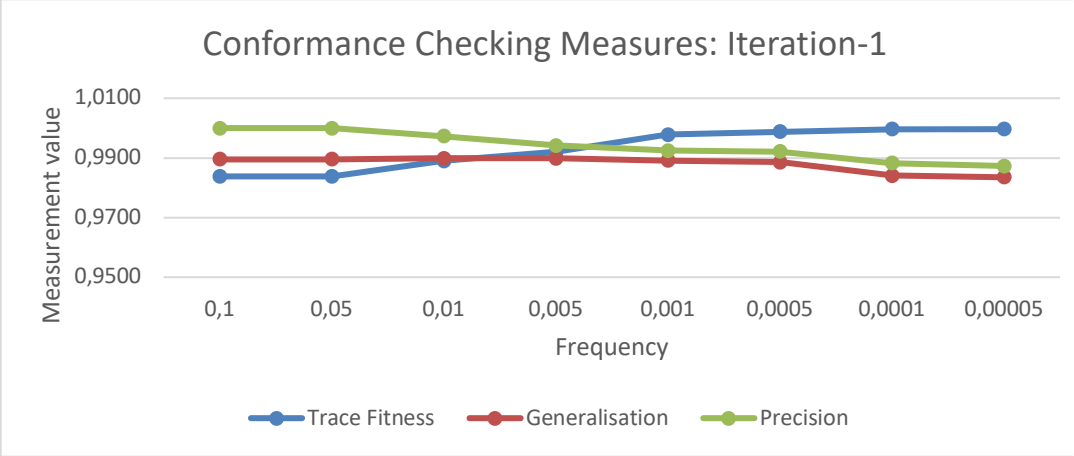


Figure 5.19 Conformance check results of the 1st iteration in Experiment 4.

Looking at the above chart of the first iteration, the next target for more sensitive adjustment was aimed at the scale of 0.005 and 0.001. The selection was made because within that scale, the trend of trace fitness and generalisation is still increasing even though the precision is declining. Furthermore, there is a cross-section between the line of trace fitness and precision between the frequency of 0.005 and 0.001.

Following the mentioned indicators, the next assumption was the optimal frequency is likely to be achieved within the frequency of 0.005 and 0.001.

The scale of the directly-follows frequency in the second iteration was changed and followed the linear pattern. The frequency values were: 0.005, 0.004, 0.003, 0.002, and 0.001. The obtained scores of trace fitness, precision, and generalisation for the second iteration are presented in Figure 5.20 below.

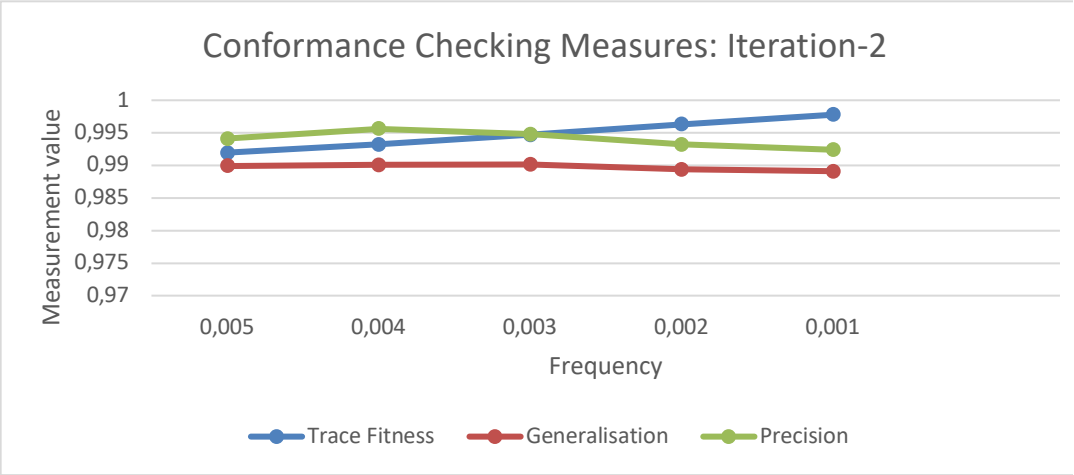


Figure 5.20 Conformance checking results of the 2nd iteration in Experiment 4.

Based on the chart in Figure 5.20, the obtained indicators were: (1) the cross-section from precision and trace fitness happened at the frequency of 0.003, (2) at the frequency of 0.004 the precision is likely had reached the optimum score and then started declining, (3) the three conformance measures are progressing positively from the frequency of 0.005 to 0.004. The later indicator was aimed to search the optimal frequency as a trade-off between the three conformance measures.

Using the above indicator, another iteration was performed for more sensitivity analysis. A linear scale was used to perform the adjustment of the directly-follow frequency from 0.005 to 0.004, the scales were 0.005, 0.0049, 0.0048, …, 0.004. The scores of trace fitness, precision, and generalisation obtained from the third iteration are presented in Figure 5.21.

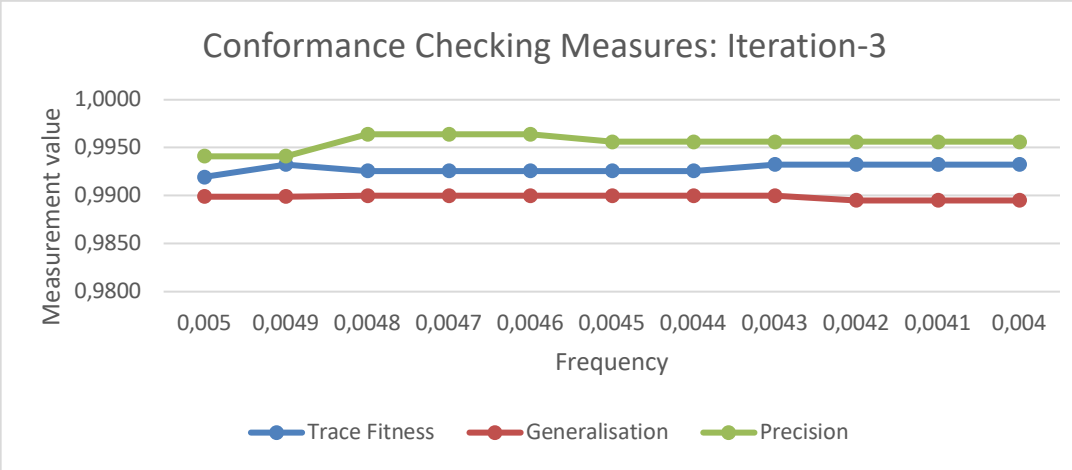


Figure 5.21 Conformance check results of the 3rd iteration in Experiment 4.

The measures from the third iteration show that the score of precision is increasing until it reaches the frequency of 0.0048 and becomes stagnant until 0.0046 and starts declining. Trace fitness was increased quite significantly at 0.0049 but dropped at 0.0048 and then began increasing slowly. Since the chart does not show any visual cues to determine the optimum frequency, the decision was made by computing the average score of the trace fitness, precision, and generalisation at each point of frequency. The average scores are presented in Table 5-32.

Table 5-32 The average score of trace fitness, precision, and generalisation.

% Directly-follow	Trace fitness	Generalisation	Precision	Average
0.005	0.9920	0.9899	0.9941	0.9920
0.0049	0.9932	0.9899	0.9941	0.9924
0.0048	0.9926	0.9900	0.9964	0.9930
0.0047	0.9926	0.9900	0.9964	0.9930
0.0046	0.9926	0.9900	0.9964	0.9930
0.0045	0.9926	0.9900	0.9956	0.9927
0.0044	0.9926	0.9900	0.9956	0.9927
0.0043	0.9932	0.9900	0.9956	0.9929
0.0042	0.9932	0.9895	0.9956	0.9928
0.0041	0.9932	0.9895	0.9956	0.9928

The table suggests that the highest average score of 0.9930 are available when the directly-follow frequencies are 0.0049, 0.0048, and 0.0047. Among those frequencies, 0.0049 was taken since it will use fewer resources when creating the disease trajectory. The lower the directly-follow frequency used – the more resources are

taken to produce a disease trajectory model. Figure 5.22 shows the thumbnail of the discovered disease trajectory model using the optimum directly-follow frequency of 0.0049 with the visualisation of a directly-follows graph (a bigger size is available at Appendix C).

5.6.2.3 Findings

The trajectory of AMI→I25-ischemic heart disease was also found in the first case study (Chapter 4). The trajectory was 410→414 using the standard code of ICD-9. The diagnostic code 410 is for acute myocardial infarction, and 414 is for ischemic heart disease. Referring to the data source, the first case study was using the MIMIC-III data set that came from a critical care unit of a private teaching hospital in Boston, USA. This current case study study-2 uses the HES-APC data set that came from the NHS hospitals in England.

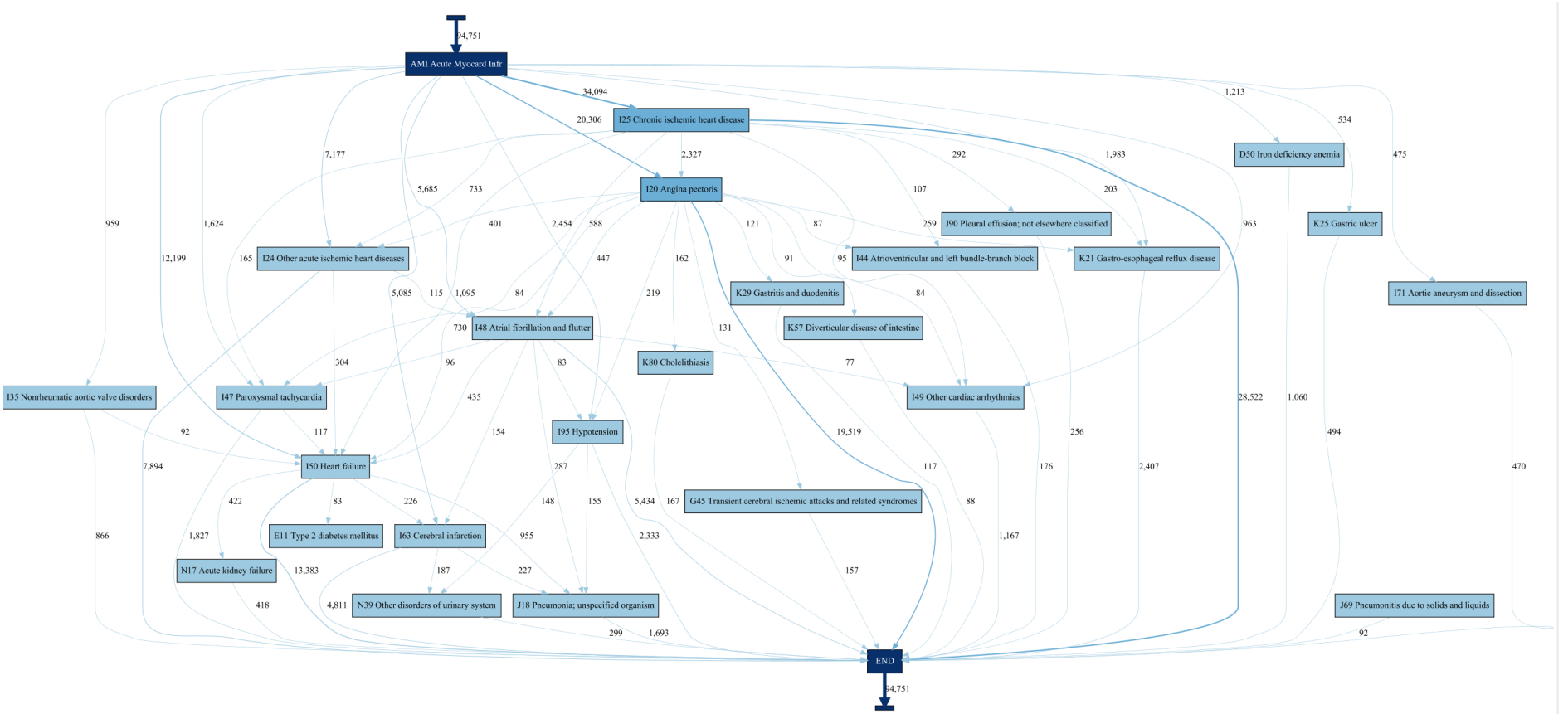


Figure 5.22 Directly-follows graph of disease trajectory model using the optimum directly-follow frequency.

5.6.3 Stage-5: Evaluation

The 5-folds cross-validation, like the previous experiments, was performed as an additional evaluation. The result is presented in **Table 5-33**.

Table 5-33 The result of 5-folds cross-validation.

#	Training folds	Validation fold	Fitness	Generalisation	Precision
1	2, 3, 4, 5	1	0.9947	0.97068	0.9719
2	1, 3, 4, 5	2	0.9948	0.9720	0.981
3	1, 2, 4, 5	3	0.9962	0.9733	0.9795
4	1, 2, 3, 5	4	0.9948	0.9763	0.9795
5	1, 2, 3, 4	5	0.9943	0.9743	0.974
Average			**0.9950**	**0.9733**	**0.9771**

The average value of each measure was expected to be lower than the measurement from the conformance checking. The event log was randomly divided into five equally sized sub-log. One sub-log was used as the test data while the other four sub-log became the training data. The lower score of the 5-folds cross-validation happened since the alignment may not properly done or there is a missing step that caused lower score. However, the average result from the 5-fold cross-validation shows that the process model has a good quality and is representative to the event log.

Chapter 6
Discussion

Chapters 4 and 5 presented the analysis of the four case studies based on the methodology in Chapter 3. The four case studies were the groundwork of the research in this thesis. In this chapter, method reflection will be discussed, together with the answers of the research questions, the challenges of using health care data, the impact of the research, and the contribution of this thesis.

6.1 Method reflection

A routinely collected electronic health record is one of the potential sources of evidence in the health care domain. Utilising the electronic health records may extract useful information that is beneficial for improving the quality of health care. In this current research, a key hypothesis was stated in Section 1.3 where *process mining is useful for identifying disease trajectories from routinely collected electronic health records.* One feasibility study in Section 3.2 followed by two case studies presented in Chapters 4 and 5 were conducted to challenge this hypothesis. In the feasibility study, process mining approach was used to recreate an existing disease trajectory model using a synthetic data set. Following the feasibility study, a process mining approach was implemented to two routinely collected electronic health records with different levels of abstraction: the EHR of a hospital and the EHR of England. This approach helped in exploring the challenges of identifying disease trajectories generally.

A standard method of conducting a process mining project, PM^2, is presented in Chapter 3. The method was initially designed for identifying business processes but was altered into a method to identify disease trajectories. Alterations were performed at two stages: the Planning stage (Stage-1) and the Data processing stage (Stage-3). At Stage-1, the *activities* of a process were translated as *diagnostic codes* and the *sequence/ trace* of these *activities* is called a *trajectory*. A significant alteration was made at Stage-3 where the event log was transformed multiple times to allow analyses based on pair of diagnostic codes.

As the two case studies used two different data sets with different levels of abstraction, then a consistent method for implementing a process mining approach is needed. The PM^2 method was chosen since it allows iterations at some stages as required and

suggests that the involvement of domain experts plays an important role in the project. Figure 6.1 presents the implementation of PM^2 as the main guideline throughout this current research. The altered stages are coloured red, while non-altered stages are coloured blue.

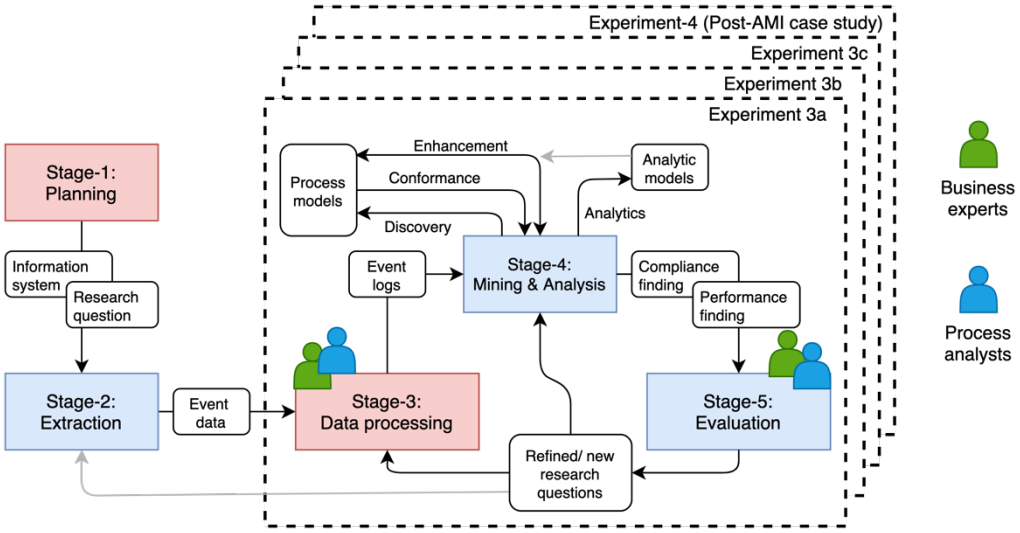


Figure 6.1 Method implementation in case study-2 based on PM^2 [10].

The iteration in PM^2 was found useful to analyse the HES-APC data set as described in the case study-2 (Section 5.6). Due to the large size of the data, Stage-1 and Stage-2 can be conducted once at the beginning of the study and the analysis was broken down into multiple experiments. The multiple experiments were conducted by using various subsets of the event data to answer the research questions. The result of the first experiment led to the next experiment and so on until all research questions are answered.

The participation of domain experts was found useful to the current research. The discussion with the domain experts since the initiation of the research had helped in designing the experiment and evaluating findings through knowledge sharing. Stage-6 of the PM^2 was not conducted unlike activities in business processes, the occurrence of diseases in disease trajectories is too difficult to be managed.

The use of PM^2 as a single method helped the author to conduct the feasibility study and the two case studies. The PM^2 framework provides a guideline to do process mining projects through stages, including the recommendation of activities, input, and the expected output.

6.2 The answers to research questions

Three research questions (RQs) presented in Section 1.3 have been broken down and answered using a feasibility study and two case studies. The aim of this current research was to provide a generalisable approach to identify disease trajectories using patients' EHR from various health care environment. The summary of all RQs being answered is presented as follows:

RQ-1: Can disease trajectories be identified using a process mining approach?

Process mining can be used to identify disease trajectories in EHR. The initial result came from a feasibility study as presented in Section 3.2. A synthetic data set was created based on a subset of a published disease trajectory model by Jensen et al. (2014) [3]. Process mining approach was successfully recreated the disease trajectory model. Following the feasibility study, in Chapter 4, the first case study was done to identify disease trajectory models in the MIMIC-III data set – an actual EHR of critical care units of a private teaching hospital in Boston, USA. The second case study presented in Chapter 5, was done by identifying disease trajectory models using process mining approaches in the HES-APC data set containing hospitalisation data of England population. The results that came from both case studies showed that disease trajectory models can be identified using process mining approaches. The extracted event data from EHR were transformed into pair logs to allow pairwise analyses to identify the disease trajectories and then retransform the pair log into an event log. Process discovery was applied to the retransformed event log to generate the disease trajectory models.

RQ-2: What are the most followed trajectories and how long were the duration of the trajectories?

In the first and second case studies, the process mining approach was able to identify and populate the disease trajectory patterns, including the frequency and the duration of each pattern, using standard and freely available process mining tools. The duration of each trajectory was calculated automatically using DISCO and available in the median and mean. The calculation of IQR that follows the median duration was done outside DISCO using a programming language. The current version of DISCO and ProM, at the time this thesis was written, did not provide IQR calculation.

The most followed trajectory patterns and median durations from the first case study are presented in Table 4-3 of Section 4.2.4. The trajectory patterns and median durations from the second case study were divided into two groups: the disease trajectories in the HES-APC data set presented in Section 5.5 (Experiment 3c) and, the disease trajectories of post-AMI incidents that presented in Section 5.6 (Experiment 4). The least common disease trajectories for post-AMI events were also available in Appendix D.

The most common disease trajectory in the HES-APC data set (Experiment 3c) was J44 COPD → J18 Pneumonia. This finding supports the study by Restrepo et al. [160] and Hartley et al [161] where reported that COPD patients are at risk of pneumonia.

RQ-3: What are the longest and shortest duration time transition trajectories?

One of the strengths of process mining is time analysis. In this current research, the identified duration of disease trajectories in DISCO was presented in the median and mean. These results can be ordered to get the longest and shortest durations. The longest and shortest time transition trajectories in case study-1 were presented in **Table 4-4** of Section 4.2.4 while the measures of case study-2 were presented in **Table 5-18**, **Table 5-30**, and **Table 5-31**.

RQ-4: Are there differences in trajectories followed by different patient groups?

There were differences in the most common disease trajectory patterns when the data were grouped by sex or age group. Differences were found in the case study-1 where the most common trajectory in the group of male patients was acute myocardial infarction followed by other forms of chronic ischemic heart disease, while in the female cohort was subarachnoid haemorrhage followed by other and ill-defined cerebrovascular disease. In the cohort of 18 to 34-year-old, the most-followed trajectory was diabetes followed by hypertensive chronic kidney disease. For the group of 35 to 64 years, the most common trajectory was acute myocardial infarction followed by ischemic heart disease while in the cohort of >64 years was acute myocardial infarction followed by heart failure.

Differences were also found in the case study-2. In Experiment 3c, the first four common trajectories were similarly occurred in both male and female cohorts, but different in the order of the frequency as presented in Table 5-19. Pattern differences of the common trajectories were also happened by age group. As presented in Table 5-20, the patient group who's younger than 35 years old don't have disease trajectories

containing AMI. The AMI began to occur in the group of patients between 35-74 years old where AMI started to be missing in the group of >74 years old.

RQ-5: What is the method to remove any repeated events and decide the direction of the trajectory?

The disease trajectories were identified using two major activities. The first was by removing any recurrence of diagnostic code and keeping the first occurrence. Second, the trajectories were determined using a method inspired by Jensen et al. (2014) and considered this method for reducing the complexity of the model as suggested in Stage 3 of the PM^2 framework. In Jensen's [3], significant pairs of diagnostic codes were identified, and each pair was combined with another pair with overlapping diagnostic codes to produce longer trajectories. The number of patients who followed the longer trajectory was then counted. This approach suggests that the trajectories of three or more diagnostic codes were composed without knowing if such trajectories are present in the data. In contrast to process mining, each patient's full diagnostic code sequence was first known and stored in the event log. This thesis proposes an enhancement to Jensen's method by applying event log transformation into pair log, filtering the less important pairs, determining the pair's directionality, and then followed by retransforming pair log into event log to form a disease trajectory model, end-to-end, using discovery algorithm.

The steps to identify disease trajectories began with creating event logs and then followed by transforming the event logs into pair logs for measuring the strength of the association of each diagnostic pair of codes. Pairs of diagnostic codes that have significantly strong associations were tested for directionality using a binomial test to get the significantly dominant trajectories. We used 50% of probability to identify which one is more dominant between $d_1 \rightarrow d_2$ or $d_2 \rightarrow d_1$ as both have the 50:50 chance. The pair log containing pairs of diagnostic codes with significantly strong association and significantly dominant trajectories were retransformed into an event log for process mining analyses including disease trajectory model discovery and conformance checking. These steps were presented in detail in 3.1 and were implemented in case study-1 (Section 4.2.3) and in case study-2 (Section 5.2.3).

RQ-6: Is the produced disease trajectory representative enough of the data? Yes.

The process mining's conformance checking helped to identify where the produced disease trajectory models in the two case studies are representative to the data. The measurement of trace fitness, generalisation, and precision show scores above 0.7 meaning that the disease trajectory models are representative of the data. Furthermore, additional evaluation using the k-fold cross-validation was done to see if the identification of disease trajectories using the process mining approach is robust. The conformance checking and the cross-validation were done every time disease trajectory discovery was conducted. The conformance checking of the first case study is presented in Section 4.2.3 and for the case study-2 are presented is Section 5.3.1.2, Section 5.5.2.2, and Section 5.6.2.2.

6.3 Challenges of using health care data

6.3.1 Data access and confidentiality

An ethical review to access and use this data set was not required. It was stated in the NHS Digital's Data Sharing Agreement (DSA) Customer Approval document number DARS-NIC-17649-G0X4B-v0.6 – section 7-Approval Consideration. However, since the data set is stored under the Secure Electronic Environment for Data (SEED), a confidentiality agreement was needed upon authorisation to access the data set, including the completion of confidentiality training course. The Information Security Awareness module which must be undertaken annually. The confidentiality agreement is part of the SEED Information Governance Policy version 3.0 dated 17 February 2014. The DSA cover letter document is presented in Appendix E.

Access to research data was disrupted during the COVID-19 outbreak. The security of SEED did not allow any connections made from outside the University network. Therefore, during the lockdown, any experiment activities that required access to the research data were postponed until further notice. The notification to regain access to the SEED was received in October 2020 (seven months after lockdown), and three months afterwards, a full access to the SEED was resolved by physically attending and using the secure University's computer network

6.3.2 Data quality

One of the challenges of working with a real-life electronic health record is the quality of the data. The EHR contains the routinely collected data during the patients' encounters to the health care service. These collected data are used for clinical

purposes and are not specifically designed for research purpose. When the data of a real-life EHR were extracted for research activity, further inspection was needed as the extraction activity was done to fulfil the data specification per request. The source of errors can be identified during the data quality inspection by following the data quality framework. The inspection activity was helpful to see if the data is suitable enough for research or to identify how much effort of cleaning and filtering is needed to make the data ready for research.

The data quality of MIMIC-III was done by Angelina Kurniati, [162] a member of the process mining research group in the University of Leeds. Kurniati summarised the quality of the data using four quality dimensions: completeness, correctness, concordance, and plausibility. A data quality framework suggested by Weiskopf & Weng [82] was used by Kurniati to assess the data quality of the MIMIC-III data set. She concluded that the MIMIC-III data set is representative to the actual EHR data with some limitations. Two major limitations are the data came from a critical unit which only contained data of patients that encountered the unit. Another limitation was no direct communication with the BIDMC hospital; learning the data was made by examining the MIMIC-III's online documentation and some related publications. Later, Kurniati concluded that the MIMIC-III data set for process mining was good quality.

The two data sets in this current study (the MIMIC-III and the HES-APC data sets) came from the actual environment of two health care organisations. The MIMIC-III data sets were extracted from the Beth Israel Deaconess Medical Center and provided by the Laboratory for Computational Physiology at the Massachusetts Institute of Technology for the purpose of research. In contrast to the MIMIC-III data set, the HES-APC data are the collection of reports from hospitals in England and were curated by the NHS Digital where they provided the extract by following the criteria as requested by the Cardiovascular Epidemiology Research Group.

According to the DARS document, the NHS Digital had filtered the data as per request, apparently the received data set still contained data that should have been excluded. For example, the data set should not contain data from patients aged less than 18 years old, but during the quality inspection their data were still exist.

6.3.3 Working with a large data set

Another challenge was the large size of the data set, it consists of 145 million rows of data and the availability of computing power to handle a large and complex data set. Due to the large size of data, it required a certain strategy to avoid consuming the computer's resources. There were two challenges in handling a large data set: first, the database server capacity to handle transaction requests, and second, the limited number of a high-specification computer to work with the data. The lesson learned from this challenge was to save every cleaning and the selection results into a hard drive to avoid high memory consumption. Once the results are saved, then the memory can be emptied for the next operation. Loading only the necessary parts of the data is required, including good record keeping of file names of every step taken in cleaning and selecting the data. One more challenge was the nature of the health care data that requires handling as a sensitive data set. It causes non-flexible working arrangements to comply with the security protocol in accessing the data.

6.3.4 Data understanding

Electronic health records in every health care management system in general was not specifically designed for research. But in the other hand, the EHR contains the digital trace of patients that is potentially valuable information to improve the quality of care and the diseases' outcome [163]. The EHR can also be seen as big data [164], and the design of every EHR may be different for every health care systems. This difference increases the challenge to understand the EHR. As a big data, EHR is defined by the five Vs: volume, variety, veracity, velocity, and value. The size of EHR (volume) can be as big as a population of a country (e.g., the HES data set of NHS England), and those data are generated and moved at a certain speed (velocity). The data may come in various forms and standards (variation) where the trustworthiness of the data (veracity) remains important for clinical and research use. Finally, the data can be useful for some domain specific (value) if being harnessed in the correct way.

The challenge in understanding the two data sets in this current research was the limited resources. Both the MIMIC-III and HES-APC data sets were studied through the online documentation that comes with the data sets and through paper publications, the direct access to the data providers was also absent. The understanding of each data set has its own challenge. The anonymisation in the

MIMIC-III data set was done by shifting the actual time into future dates, and the number of years in the actual time window (between 2001 and 2012) is less than the future time window (between 2100 and 2200). This anonymisation prevented the analysis of the burden of disease over time to the health care provider. The challenge of analysing the MIMIC-III is presented in Section 4.1.

A similar reason to the MIMIC-III data set, to prevent reidentification to the patients, the timestamp data in the HES-APC data set were provided in months and years. The incomplete timestamps prevented the accuracy of the disease trajectories' duration and were presented in months instead. The reliability of comparing diseases in trajectories between the MIMIC-III and the HES-APC data sets was low as the coding standard of disease in both data set is different. The diseases in MIMIC-III were coded using the ICD-9, while the HES-APC data set used the newer version, the ICD-10.

6.4 Impact of disease trajectory mining in health care

Mining disease trajectories using routinely collected data can bring benefits to many different perspectives. From the perspective of health services, hospitals can utilise their data even more to get more information on how the diseases are progressed. Clinicians can be reinforced with the information of the patient's disease trajectory to see the "footprints" of diseases and how the diseases had progressed during their patient's lifetime. This information can help the clinicians to create better intervention plan for the patients, including what would be the potential burden. The mined common pattern of disease trajectories extracted from one health care provider's EHR may differ from another health care provider. The differences can provide support to help the health care provider in creating a plan for improvement. As a first step, using the large scale of data from the NHS, we can provide information on the burden of disease.

From the epidemiology perspective, mining disease trajectories enables to quantify the long-term burden of health care providers, e.g., supporting the actual quantification of epidemiology to predict the occurrence of diseases. The presented method in this thesis is also useful for hypothesis generation that supports further research questions.

6.5 Contribution of this thesis

To contribute to the process mining community and the health care research was the motivation behind this research. There are three main contributions of this thesis: 1) a method to identify disease trajectories from electronic health records using process mining approaches, 2) a feasibility study and case studies of mining disease trajectories using process mining, and 3) a tool written in Python for identifying disease trajectories using event log extracted from EHR. Each of the contributions is described in the following sections, and the extension of the work in this thesis by future process mining researchers is expected.

6.5.1 A method to identify disease trajectories from EHR

The first contribution of this study was the method of identifying disease trajectories. The method was applied in a case study and resulted in a jointly authored publication which was presented at a virtual event of the PODS4H 2020. The publication presents the use of process mining approach to identify disease trajectories from the MIMIC-III data set, as described in Chapter 4.

The method was conducted by following a process mining methodology, the PM^2. The sequence of diseases was viewed as a process since it has the properties for process mining analysis: case id, activity, and time stamp. The patient id was chosen as the case id, the standard ICD code of disease was used as the activity name, and the time when the disease occurred was used as the time stamp. Event log transformations were added under the Filtering Log activity as part of the Stage-3 - Data processing. The event log transformations are needed for filtering event logs by removing the non-important pairs of diagnostic codes. The method of identifying disease trajectories by Jensen et al. (2014) was used as a filtering strategy to reduce the complexity of the data and focus on the important disease trajectories. Using the process mining approach, a disease trajectory model was able to be discovered using process mining tools of ProM Framework (https://www.promtools.org) and DISCO (https://fluxicon.com), which was difficult to do with only using the Jensen et al.'s method. Conformance checking was also playing an important role in measuring the quality of the discovered disease trajectory model, as this measurement is missing in Jensen et al.'s method. The quality of the discovered disease trajectory model was measured to determine whether the model was representative of the data or vice versa.

The process mining approach of identifying disease trajectories was found to be robust. The conformance checking results were evaluated using the 5-folds cross-validation, where each validation resulting high values of replay fitness, precision, and generalisation.

6.5.2 Assessing the application of process mining using EHRs

The first assessment of applying process mining to identify disease trajectory was the feasibility study, as presented in Section 3.2. The study was conducted to quantify that process mining approach can be used to mine disease trajectories. The feasibility study was also used to explore how process mining approach and tools such as ProM and DISCO can be implemented while the main purpose of process mining is for analysing data from the perspective of process. The idea behind process mining is to extract knowledge from event logs containing records of actual processes that are available in the information system. On the other hand, a disease is not a process by definition. It has the property of process mining problems when the diseases occur as a sequence over time. The time facilitates the order of the sequence of diseases, and the sequence is identified using the patient id.

Following the result of the feasibility study, the method was used in two case studies using two different sets of data with two different abstractions. The first case study was done using a hospital-level EHR – the MIMIC-III data set and the second case study used a nation-wide EHR – the NHS England's HES-APC data set.

The first case study, as presented in Chapter 4, shows that disease trajectories can be mined from the MIMIC-III data set. A jointly-authored publication was produced and presented at an international virtual conference of the Process-oriented Data Science for Health (PODS4H) 2020 [86]. Disease trajectories can be mined from the MIMIC-III data set and the results support the previous finding in term of disease comorbidities.

The second case study had been presented in Chapter 5 using the HES-APC data set. Due to the size of the data and the access restriction, working with the HES-APC data set became a challenging process. The data set should remain within a secure environment and can only be accessed from within the University using a secure private network. A special high-specification computer facility was used to perform

process mining analyses using the ProM tools and DISCO. To understand the data, online resources were learned including the official documentations provided by the NHS Digital, previous publications that used the HES-APC data set, and through a series of discussions with the clinical experts. The second case study showed that the process mining method can be conducted using two different samples approach of the cohort.

6.5.3 Tool for identifying disease trajectories

The third contribution of this study is a tool written in Python for identifying disease trajectories. The tool could help in preparing event log that contains event data of disease trajectories extracted from electronic health records. The input of this tool is a collection of event data that contains patient identifiers, their respective primary diagnostic codes, and time stamps for every diagnostic codes' occurrence. The input data are the clean version of event data that has been through a data quality evaluation and a series of inclusion or exclusion criteria as required for answering the research questions.

The tool processes the input data by sorting the event log, making selection on the first ever occurrence of each diagnostic codes for each patient, making selection on any interests of specific diagnostic codes, transforming event log into pair log, filtering the pair log and select the strongly associated and significant pairs of diagnostic codes, retransforming the pair log into an event log, and enriching the event log. The output of this tool is an event log ready for loading into process mining tools for analyses. The tool for disease trajectories is available on GitHub: https://github.com/gpkusuma/distories.

Chapter 7
Summary

Chapter 4 and Chapter 5 presented the analysis of the three studies underpinning the research in this thesis. These chapters were structured using the method described in Chapter 3. In this final chapter, the work is summarised before reflections are made on the methods and the findings from each study. This chapter concludes by assessing whether the hypothesis and primary research questions have been effectively addressed and answered, before discussing the contributions of this thesis, its impact, and considerations for future directions.

7.1 Conclusions

This section covers the conclusions from the literature review, developed methodology, experiment using the MIMIC-III data set, experiment using the HES-APC data set, and discussion.

7.1.1 Conclusion from the literature review

A process mining of disease trajectory literature review was performed during the study. Provided below are the important finding regarding the literatures:

1. To date, there are a very limited number of process mining research for identifying disease trajectories and studies to identify general disease trajectories at national level are still vacant. These finding suggest opportunities in the domain.
2. Previous studies mostly used specific cohort e.g., patients with Type-2 Diabetes or sepsis. A general disease trajectory study is open for further exploration.
3. There are three approaches to define the trajectory: first, using some statistical analysis of correlation measurement and a binomial test; second, using n-grams; and the latter by utilising a process mining algorithm. In addition, a recently published literature suggests the utilisation of disease protein to establish the trajectory among diseases. Disease protein is a type of disease where the structure of some proteins become abnormal and disrupts the cells' function, body tissues and organs. The protein-to-protein interaction performs

its function for the human body. The network of interactions represents their associations, which significantly impact human health.

4. The implementation of process mining of disease trajectories was found in three ways: process discovery, conformance checking, and for model visualisation.

7.1.2 Conclusion from the methodology development

The steps to perform disease trajectory mining were built based on the process mining framework PM^2 with the addition of a question-driven methodology. The crucial aspect of performing disease trajectory analysis using process mining is the collaboration with experts with medical knowledge. The study should build on health informatics, medical informatics, process mining or data science in general with the involvement of experts since the beginning of the study.

The main part of this study is the data processing stage (Stage-3). The event log filtering, the transformation of the event log into a pair log, the statistical analyses, and the retransformation of a pair log into an event log. Depending on the research question(s), the event log filtering is to include or exclude the less important information or vice versa, for further analysis. The event log transformation into a pair log is to enable the statistical analyses since the analyses require pairs of diagnostic codes. The statistical analyses perform the measurement association between diagnostic codes and the directionality test. Finally, the retransformation of pair log into event log for process mining analysis using process mining tools.

7.1.3 Conclusion from the experiment on the MIMIC-III data set

Disease trajectory was modelled from the MIMIC-III data set using process mining techniques. The results of the study showed how process mining techniques could be used to quickly produce disease trajectory models from event log data. Two process mining tools were used for two purposes; the DISCO was used to produce disease trajectory models and calculated the time durations of each trajectory, where the ProM was used to conduct conformance checking besides discovering disease trajectory models. However, within the two process mining tools, the inter-quartile range (IQR) of median duration was not available and needed to be calculated elsewhere. In this study, the IQRs were calculated using Python.

The disease trajectories presented above have been created using the DISCO process mining tools and the MIMIC-III data described in Chapter 4. Here the median frequency of each trajectory is displayed but not the IQR.

This work is the extension of previous work by a member of the University of Leeds process mining research group, Samantha Sykes. Sykes had shown the opportunity to use process mining techniques for modelling disease trajectory. Her method could replicate Jensen et al.'s method, although there were some limitations in retaining patient information. The extended work presented in this thesis could retain the patient information so the trajectories of every individual can be followed. A feasibility study to prove the patient information can be retained was conducted and a jointly authored publication with Sykes and other contributors was produced.

The limitations discussed in these sections include the method in the case study-1 where the association of diagnostic pairs was taken based on relative risk measurement without the null hypothesis significance testing. The method was improved in the case study where null hypothesis significance-testing was conducted. The second limitation is the exclusion of the secondary diagnoses may contain existing comorbidities, which have the potential for a better understanding of the disease. The third limitation is the deidentification of the MIMIC-III data set, where one of the steps was by shifting the actual time into the future time. The shifting prevented the identification of the most burdened times.

The next limitation is the representativeness of the data set. MIMIC-III data set was collected from a private hospital in Boston, US, therefore it doesn't represent the population of Boston. The health system in the US allows people to get medical services from private health care providers. Had the data collected in the MIMIC-III data set, it is possible that the data of the same patients exist in other health care providers which may contain important information for disease trajectory modelling.

7.1.4 Conclusion from the experiment on the HES-APC data set

This study successfully mined disease trajectory models for the HES-APC data set using the process mining techniques. The HES-APC data set used in this study was appropriate for this study and was made available for reproducibility. The HES-APC data set is available for research which can be requested through the NHS Digital. Properties to create an event log for process mining analysis are available in the data set. The data set was pseudonymised and available for research access through NHS

Digital's Data Access Request Service (DARS) Application together with Data Sharing Framework Contract (DSFC). One benefit of using the HES-APC data set is the availability for many different types of analyses.

The obtained HES-APC data set contains patients' hospitalisation data from 2008 until 2017. With the length of the data, the identification of trajectories with two or more diagnostic codes was able to be conducted. Furthermore, with the rich variation of diseases, mining disease trajectories of patients with a specific condition were able to be done. The last benefit was described in Section 5.6 where patients who had diagnosed with AMI were analysed.

The challenge of using the HES-APC data set was due to the large size of the data set and the security procedure to access the data set. The obtained data set has the size of 145 million records and took around 115 gigabytes of storage space. A secure connection from within the University network and a high-specification computer were used to access the data. The limitation of the server's transactional memory had increased the level of complexity to process the data.

The limitation of using the HES-APC data set was due to the pseudonymisation of the data. The timestamps were available for months and years without the presence of the dates. This limitation prevented to provide a more precise duration of the trajectories.

7.1.5 Conclusion from the discussion

This study identifies disease trajectory models using process mining techniques – a process-oriented data analytics approach. This study used actual EHRs that are already available in the health care information system and a synthetic data set to test the feasibility of the method. The data set contains patient records during the encounter with the health care provider that are routinely collected to support clinicians or for administrative purposes. The common limitation of EHR is data quality where many factors could contribute to the quality of the EHR, starting from how the data is being acquired, recorded, or managed, let alone when the health care provider must adopt some changes due to new standards, regulations, or internal development.

Other data challenges of process mining in health care include data access, ethics, and data understanding. Health care data contain records about patient information and other sensitive information. Therefore, the data needs to be protected and handled carefully. The protections come in various ways, from formal regulations to technical

aspects. Working with such data requires an adequate understanding of data protection and ethics.

Another crucial aspect is data understanding, where information regarding the data can only be retrieved from the third-party sources. Direct communication with those who are directly working with the data is not available.

7.2 Presentation and feedback

The works presented in this thesis are conducted as contributions to the community of the health care process mining including the Process-Oriented Data Science for Health (PODS4H). Extended communities of associated study may also included for contributions. The three contributions of this thesis are the case studies of mining disease trajectories for health care process mining, the transformation of event logs into pair logs, and the event log generator for disease trajectory analysis. This study contributes to the technical aspects and hypothesis generation of the health care process mining domain.

This study produced seven published contributions, there were three poster presentations, three papers presented at international conferences, and one paper published in an international journal, as presented at the beginning of this thesis. Some parts of this study were presented in four events outside of the above. Those events are:

1. School of Computing Ph.D. Symposium (11 April 2018)
   The Ph.D. symposium is an annual event being held exclusively in the School of Computing, University of Leeds. The event was attended by around 20 postgraduate researchers doing15 minutes presentations of their research. The audience is the School of Computing staff, including the postgraduate researchers and students. The presentation started with an introduction and then followed by the presentation of my research plan i.e., the background, hypothesis, research questions and a brief method to conduct the research.
2. Farr Institute Symposium and Innovation Workshop (21-22 May 2018)
   The event was held by the Farr Institute, a national consortium in data science for health. The event participants were PhD students from across the UK at any stage of their study. All of them were conducting research using data to answer research questions in health care. Each participant presented the

ongoing research or preliminary work, including the current challenge in any aspect of the study.

The feasibility study in Section 3.2 of this thesis was presented at the event. The most interesting discussion was the challenge of working using data from the NHS England.

3. Colloquium of the Applied Information System research group at Telkom University.

   A colloquium is a monthly event in the Applied Information System research group. The audience was a member of the research group, and they had various research interests. The main intention was to introduce the process mining technique to the research group, including a case study from this thesis and other case studies from the members of the University of Leeds's process mining research group (all case studies were presented with permission).

4. LIDA Seminar Series of the University of Leeds (6 February 2020)

   The seminar series is a monthly event being held by the Leeds Institute of Data Analytics of the University of Leeds. The audience was mostly health care professionals, IT professionals from health care industries, and Ph.D. students who have an interest in health informatics.

   The most interesting discussion was the implementation of mining disease trajectories using process mining on a specific cohort who had a similar underlying health condition. This topic has been done to identify disease trajectory models from a group of patients who had an acute myocardial infarction. The work is presented in Section 5.6.

5. Webinar Series of School of Applied Science Telkom University (21 July 2020)

   This online seminar was held by the School of Applied Science at Telkom University. The audience was mostly vocational students and the lecturers of information system program, and the members of applied information system research group.

   The presentation contains an introduction to data science in health care, the emerging process mining techniques and showcasing the example of process mining implementation in the domain of health care.

6. Global Learning Week 2021 (GLOW 2021) of Telkom University.

This is an annual fortnight event since 2015 for the undergraduate international students in the School of Computing, Telkom University. The students were in their third year and some of them were in their fourth year of their study. Among those audiences, there were invited master students who have an interest in process mining. The main lecturer was Owen Johnson from the University of Leeds who was assisted by two lecturers from the School of Computing Telkom University.

Some parts of the work in this thesis were presented at the event, together with other invited speakers – the members and alumni of the process mining health care research group at the University of Leeds.

7.3 Future work

This study has the potential for implementation using a different data set. Moreover, it can be improved in future work. The improvement to this study can be made in five courses, and the descriptions are provided as follows.

The first course is to increase the detail of the disease by incorporating the lower abstraction of diagnostic code. The works in this thesis use the first three characters of the ICD-9 and ICD-10 coding standards as the category of diagnostic codes. The maximum length of diagnostic codes for recording is up to five characters for the ICD-9 code and up to seven characters for the ICD-10 code. The additional fourth or fifth characters in both coding standards are used to mark subcategories to provide more detailed information [106, 165]. For example, the diagnostic code for acute myocardial infarction in the ICD-9 code is 410 or I21 in the ICD-10 code. Adding zero (“0”) as the fourth characters into the codes above, it gives more detail information: 410.0 is acute myocardial infarction of the anterolateral wall, while I21.0 means acute myocardial infarction of another anterior wall. Using more detailed diagnostic codes will improve the accuracy of the trajectory models but also increase the requirement of computing power as the size of the event log is increasing.

Another way to increase the accuracy of the trajectory models is by incorporating secondary diagnostic codes. This second course will include other diagnoses, symptoms, underlying conditions, problems, complaints, or any other reasons that came together with the primary diagnostic code. The inclusion of secondary diagnostic codes may add more detailed trajectory models.

The third course would be conducting a case-control study to identify possible risk factors. The pattern of disease trajectory models of a group of patients with a certain condition is compared to those who never have the condition. This type of study could be useful in the domain of epidemiology.

Another type of comparison in the fourth course is by comparing disease trajectory models based on the variation of the patients' location. It could be a comparison of the characteristics of two disease trajectory models from two different health care providers, regions, or even countries.

The fifth course of future work is to study prediction. Using the patient's current disease trajectory, the study will identify if there is/ are any disease trajectory model(s) that are relatively similar to the patient's characteristics. If the identification is successful, then it is possible to predict the next possible diagnostic codes or outcomes. This could be useful to inform the clinicians for designing an intervention for the patients.

Besides the above improvements, process mining, in general, could benefit clinical practice. A range of techniques in process mining can be used to improve the quality of treatment by analysing positive deviations or reducing variance in the treatment or intervention procedure. Process mining benefits the clinical practice by improving the overall quality of the organisational business processes. Non-clinical and clinical support are two services that can be improved. These services probably influence the perception of the patients and their families about the practice's overall service quality." [166].

7.4 Final remark

Process mining has been implemented in the health care domain, mainly for answering problems or research questions with the orientation of process. This thesis has shown that process mining approaches can be used to identify disease trajectory models – the non-process-oriented problems. At least three characteristics must be fulfilled for a problem to be seen as a process mining problem. From here an event log can be produced as the input of process mining. The three characteristics are: a sequence, event, and instance or case to mark a unique sequence of events.

The proposed method in this study has been implemented to mine disease trajectory models from three different electronic data sets. Each data set represents a different level of settings, from a small synthetic EHR, an anonymised actual hospital level

EHR, and an anonymised hospitalisation EHR of a population. Using process mining approach, the disease trajectories characteristics of patients of an entity can be identified, and compared to see the differences between the same level of the entity. This thesis is focused on the usage of the already available process mining tools for quickly generating disease trajectories.

In conclusion, this thesis has shown the potential of process mining in identifying disease trajectories and provides valuable insights into the application of process mining in the health care domain. The results of this study highlight the feasibility of using process mining techniques to mine disease trajectory models and demonstrate the usefulness of the approach in characterizing the disease progression of patients. The study also highlights the importance of having a well-structured event log as input to the process mining techniques and highlights the need for further research in this area to better understand the full potential of process mining in the healthcare domain. Overall, this thesis makes a significant contribution to the field of process mining and its application in healthcare, and lays the foundation for further research in this area.

List of References

1. Boyd, K.M. Disease, illness, sickness, health, healing and wholeness: Exploring some elusive concepts. *Medical Humanities.* 2000, **26**(1), pp.9-17.

2. Hidalgo, C.A., Blumm, N., Barabási, A.-L. and Christakis, N.A. A Dynamic Network Approach for the Study of Human Phenotypes. *PLoS Computational Biology.* 2009, **5**(4), pp.e1000353-e1000353.

3. Jensen, A.B., Moseley, P.L., Oprea, T.I., Ellesøe, S.G., Eriksson, R., Schmock, H., Jensen, P.B., Jensen, L.J. and Brunak, S. Temporal disease trajectories condensed from population-wide registry data covering 6.2 million patients. *Nature Comm.* 2014, **5**(May), pp.1-10.

4. Roque, F.S., Jensen, P.B., Schmock, H., Dalgaard, M., Andreatta, M., Hansen, T., Søeby, K., Bredkjær, S., Juul, A., Werge, T., Jensen, L.J. and Brunak, S. Using Electronic Patient Records to Discover Disease Correlations and Stratify Patient Cohorts. *PLoS Computational Biology.* 2011, **7**(8), pp.e1002141-e1002141.

5. Westergaard, D., Moseley, P., Sørup, F.K.H., Baldi, P. and Brunak, S. Population-wide analysis of differences in disease progression patterns in men and women. *Nature Communications.* 2019, **10**(1), pp.666-666.

6. van der Aalst, W.M.P. *Process Mining: Data Science in Action.* 2 ed. Springer-Verlag Berlin Heidelberg, 2016.

7. Chandy, K.M., Charpentier, M. and Capponi, A. Towards a theory of events. *ACM International Conference Proceeding Series.* 2007, **233**(May 2014), pp.180-187.

8. World Health Organization. *International Statistical Classification of Diseases and Related Health Problem (ICD).* [Online]. 2020. [Accessed]. Available from: https://www.who.int/standards/classifications/classification-of-diseases

9. van der Aalst, W.M.P. *Process Mining: Discovery, Conformance and Enhancement of Business Processes.* Berlin, Heidelberg: Springer, 2011.

10. van Eck, M.L., Lu, X., Leemans, S.J.J. and van Der Aalst, W.M.P. PM2: A process mining project methodology. In: Zdravkovic, J.Kirikova, M.Johannesson, P. eds. Springer, Cham, 2015, pp.297-313.

11. Beth Israel Deaconnes Medical Center. *About BIDMC.* [Online]. 2020. [Accessed]. Available from: https://www.bidmc.org

12. The U.S. Census Bureau. *QuickFacts Boston city, Massachusetts; United Statess.* [Online]. 2020. [Accessed]. Available from: https://www.census.gov/quickfacts/fact/table/bostoncitymassachusetts,US/PST045219

13. Office for National Statistics. *England population mid-year estimate.* [Online]. 2020. [Accessed]. Available from: https://www.ons.gov.uk

14. Johnson, A.E.W., Pollard, T.J., Shen, L., Lehman, L.W.H., Feng, M., Ghassemi, M., Moody, B., Szolovits, P., Anthony Celi, L. and Mark, R.G. MIMIC-III, a freely accessible critical care database. *Scientific data.* 2016, **3**, p.160035.

15. Fessenden, M. Protein maps chart the causes of disease. *Nature.* 2017, **549**(7671), pp.293-295.

16. Hill, A.B. The environment and disease: association or causation? *Journal of the Royal Society of Medicine.* 2015, **108**(1), pp.32-37.

17. American Medical Association. *Defining Basic Health Care - Code of Medical Ethics Opinion.* [Online]. 2020. [Accessed 26-March-2020]. Available from: https://www.ama-assn.org

18. World Health Organization: Strengthening health systems to improve health outcomes: WHO's framework for action. *Geneva: WHO.* 2007.

19. Hood, L., Heath, J.R., Phelps, M.E. and Lin, B. Systems Biology and New Technologies Enable Predictive and Preventative Medicine. *Science.* 2004, **306**(5696), pp.640-643.

20. Flores, M., Glusman, G., Brogaard, K., Price, N.D. and Hood, L. P4 medicine: how systems medicine will transform the healthcare sector and society. *Personalized Medicine.* 2013, **10**(6), pp.565-576.

21. McGinnis, J.M., Powers, B. and Grossmann, C. Digital infrastructure for the learning health system: the foundation for continuous improvement in health and health care: workshop series summary. 2011.

22. Atta, H.M. Edwin smith surgical papyrus: The oldest known surgical treatise. *American Surgeon.* 1999, **65**(12), pp.1190-1192.

23. Arab, S.M. Medicine in Ancient Egypt Part 1 of 3. *Arab World Books.* 2017.

24. Hemingway, H., Asselbergs, F.W., Danesh, J., Dobson, R., Maniadakis, N., Maggioni, A., Van Thiel, G.J.M., Cronin, M., Brobert, G., Vardas, P., Anker, S.D., Grobbee, D.E. and Denaxas, S. Big data from electronic health records for early and late translational cardiovascular research: Challenges and potential. *European Heart Journal.* 2018, **39**(16), pp.1481-1495.

25. World Health Organisation. *International Classification of Diseases Eleventh Revision (ICD-11)*. 2022.

26. Pinaire, J., Chabert, E., Azé, J., Bringay, S. and Landais, P. Sequential Pattern Mining to Predict Medical In-Hospital Mortality from Administrative Data: Application to Acute Coronary Syndrome. *Journal of Healthcare Engineering.* 2021, **2021**, p.5531807.

27. Kusuma, G., Sykes, S., McInerney, C. and Johnson, O. Process Mining of Disease Trajectories: A Feasibility Study. In: *13th International Conference on Health Informatics.* 2020, pp.705-712.

28. van der Aalst, W.M.P. and Stahl, C. *Modeling Business Processes: A Petri Net-Oriented Approach.* The MIT Press, 2011.

29. Lang, M., Bürkle, T., Laumann, S. and Prokosch, H.-U. Process Mining for Clinical Workflows: Challenges and Current Limitations. In: *eHealth Beyond the Horizon – Get IT There*, *2008*, 2008, pp.229-234.

30. Aalst, W.v.d., Weijters, T. and Maruster, L. Workflow mining: discovering process models from event logs. *IEEE Transactions on Knowledge and Data Engineering.* 2004, **16**(9), pp.1128-1142.

31. de Medeiros, A.K.A., van der Aalst, W.M.P. and Weijters, A.J.M.M. Workflow Mining: Current Status and Future Directions. In: *Berlin, Heidelberg*. Springer Berlin Heidelberg, 2003, pp.389-406.

32. Günther, C.W. and van der Aalst, W.M.P. Fuzzy Mining – Adaptive Process Simplification Based on Multi-perspective Metrics. In: Berlin, Heidelberg: Springer, Berlin, Heidelberg, 2007, pp.328-343.

33. Leemans, S.J.J., Fahland, D. and Van Der Aalst, W.M.P. Discovering block-structured process models from event logs - A constructive approach. In: *2013*: Springer, Berlin, Heidelberg, pp.311-329.

34. Weijters, A.J.M.M., Van Der Aalst, W.M.P. and Alves De Medeiros, A.K. Process Mining with the HeuristicsMiner Algorithm. In: Eindhoven University of Technology, The Netherlands, 2006.

35. Zhang, X. and Chen, S. Pathway identification via process mining for patients with multiple conditions. In: *2012*.

36. Buijs, J.C.A.M., Van Dongen, B.F. and Van Der Aalst, W.M.P. On the Role of Fitness, Precision, Generalization and Simplicity in Process Discovery. In: *On the Move to Meaningful Internet Systems: OTM 2012*. Springer, Berlin, Heidelberg, 2012.

37. Sarker, I.H., Alqahtani, H., Alsolami, F., Khan, A.I., Abushark, Y.B. and Siddiqui, M.K. Context pre-modeling: an empirical analysis for classification based user-centric context-aware predictive modeling. *Journal of Big Data.* 2020, **7**(1), pp.1-23.

38. Valdés, J.J., Céspedes-González, Y. and Pou, A. Process Mining as a Time Series Analysis Tool via Conformance Checking. In: *Cham*. Springer International Publishing, 2022, pp.636-649.

39. Janssenswillen, G., Donders, N., Jouck, T. and Depaire, B. A comparative study of existing quality measures for process discovery. *Information Systems.* 2017, **71**, pp.1-15.

40. Van der Aalst, W., Adriansyah, A. and Van Dongen, B. Replaying history on process models for conformance checking and performance analysis. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery.* 2012, **2**(2), pp.182-192.

41. Greco, G., Guzzo, A., Pontieri, L. and Sacca, D. Discovering expressive process models by clustering log traces. *IEEE Transactions on Knowledge and Data Engineering.* 2006, **18**(8), pp.1010-1027.

42. Rozinat, A. and van der Aalst, W.M.P. Conformance checking of processes based on monitoring real behavior. *Information Systems.* 2008, **33**(1), pp.64-95.

43. Adriansyah, A., Munoz-Gama, J., Carmona, J., van Dongen, B.F. and van der Aalst, W.M.P. Alignment Based Precision Checking. In: *Business Process Management Workshops*, *2013//*, *Berlin, Heidelberg*. Springer Berlin Heidelberg, 2013, pp.137-149.

44. Mendling, J., Neumann, G. and van der Aalst, W. Understanding the Occurrence of Errors in Process Models Based on Metrics. In: *Berlin, Heidelberg*. Springer Berlin Heidelberg, 2007, pp.113-130.

45. Yasmin, F., Bukhsh, F. and Silva, P. *Process Enhancement in Process Mining: A Literature Review.* 2018.

46. van der Aalst, W.M.P., Adriansyah, A., de Medeiros, A.K.A., Arcieri, F., Baier, T., Blickle, T., Bose, J.C., van den Brand, P., Brandtjen, R., Buijs, J., Burattin, A., Carmona, J., Castellanos, M., Claes, J., Cook, J., Costantini, N., Curbera, F., Damiani, E., de Leoni, M., Delias, P., van Dongen, B.F., Dumas, M., Dustdar, S., Fahland, D., Ferreira, D.R., Gaaloul, W., van Geffen, F., Goel, S., Günther, C., Guzzo, A., Harmon, P., ter Hofstede, A., Hoogland, J., Ingvaldsen, J.E., Kato, K., Kuhn, R., Kumar, A., La Rosa, M., Maggi, F., Malerba, D., Mans, R.S., Manuel, A., McCreesh, M., Mello, P., Mendling, J., Montali, M., Motahari-Nezhad, H.R., zur Muehlen, M., Munoz-Gama, J., Pontieri, L., Ribeiro, J., Rozinat, A., Seguel Pérez, H., Seguel Pérez, R., Sepúlveda, M., Sinur, J., Soffer, P., Song, M., Sperduti, A., Stilo, G., Stoel, C., Swenson, K., Talamo, M., Tan, W., Turner, C., Vanthienen, J., Varvaressos, G., Verbeek, E., Verdonk, M., Vigo, R., Wang, J., Weber, B., Weidlich, M., Weijters, T., Wen, L., Westergaard, M. and Wynn, M. Process Mining Manifesto. In: Springer, Berlin, Heidelberg, 2012, pp.169-194.

47. Johnson, O.A., Ba Dhafari, T., Kurniati, A., Fox, F. and Rojas, E. The ClearPath Method for Care Pathway Process Mining and Simulation. In: *Business Process Management Workshops. BPM 2018.*: Springer, Cham, pp.239-250.

48. Bozkaya, M., Gabriels, J. and Werf, J.M.v.d. Process Diagnostics: A Method Based on Process Mining. In: *2009 International Conference on Information, Process, and Knowledge Management*, *1-7 Feb. 2009*: IEEE Press, 2009, pp.22-27.

49. Rebuge, Á. and Ferreira, D.R. Business process analysis in healthcare environments: A methodology based on process mining. *Information Systems.* 2012, **37**(2), pp.99-116.

50. Rojas, E., Sepúlveda, M., Munoz-Gama, J., Capurro, D., Traver, V. and Fernandez-Llatas, C. Question-Driven Methodology for Analyzing Emergency

Room Processes Using Process Mining. *Applied Sciences.* 2017, **7**(3), pp.302-302.

51. Diba, K. Towards a comprehensive methodology for process mining. In: *Proceedings of the 11th Central European Workshop on Services and their Composition, Bayreuth*, 2019, pp.9-12.

52. Process Mining Group. *ProM - Process Mining Toolkit.* [Online]. 2010. [Accessed]. Available from: https://www.promtools.org

53. Verbeek, H.M.W., Buijs, J.C.A.M., van Dongen, B.F. and van der Aalst, W.M.P. ProM 6: The Process Mining Toolkit. *Business Process Manahement.* 2010.

54. Verbeek, H.M.W., Buijs, J.C.A.M., van Dongen, B.F. and van der Aalst, W.M.P. XES, XESame, and ProM 6. In: *Berlin, Heidelberg*. Springer Berlin Heidelberg, 2011, pp.60-75.

55. Günther, C.W. and Rozinat, A. Disco: discover your processes. In: *Demonstration Track of the 10th International Conference on Business Process Management, BPM Demos 2012*: CEUR-WS. org, 2012, pp.40-44.

56. Celonis Gmb, H. *Celonis.* [Online]. 2019. [Accessed]. Available from: https://www.celonis.com/

57. Yang, W. and Su, Q. Process Mining for Clinical Pathway Literature Review and Future Directions. *Service Systems and Service Management (ICSSSM), 2014 11th International Conference.* 2014, (2010), pp.1-5.

58. Rojas, E., Arias, M. and Sepúlveda, M. Clinical Processes and Its Data, What Can We Do with Them? In: *2015*: SCITEPRESS - Science and and Technology Publications, 2015, pp.642-647.

59. Rojas, E., Munoz-Gama, J., Sepulveda, M. and Capurro, D. Process mining in healthcare: A literature review. *Journal of Biomedical Informatics.* 2016, **61**(April), pp.224-236.

60. Erdoğan, T. and Tarhan, A. Process Mining for Healthcare Process Analytics. In: *2016*, *Berlin*. Conference Publishing Services, 2016, pp.125-130.

61. Erdogan, T.G. and Tarhan, A. Systematic Mapping of Process Mining Studies in Healthcare. *IEEE Access.* 2018, **6**, pp.24543-24567.

62. Batista, E. and Solanas, A. Process mining in healthcare: A systematic review. In: *9th International Conference on Information, Intelligence, Systems and Applications (IISA)*, *2018*, *Zakynthos, Greece*. IEEE, 2018.

63. Mannhardt, F. and Blinde, D. *Analyzing the trajectories of patients with sepsis using process mining.* APA, 2017.

64. Hendricks, R. Process Mining of Incoming Patients with Sepsis. *Online Journal of Public Health Informatics.* 2019, **11**(2).

65. Kurniati, A.P., Hall, G., Hogg, D. and Johnson, O. Process mining in oncology using the MIMIC-III dataset. In: *2018/03//*, pp.12008-12008.

66. Farid, N., De Kamps, M. and Johnson, O. Process Mining in Frail Elderly Care: A Literature Review. In: *2019*: SciTePress, Science and Technology Publications, 2019, pp.332-339.

67. Williams, R., Rojas, E., Peek, N. and Johnson, O.A. Process mining in primary care: A literature review. In: *Studies in health technology and informatics*, *2018*, pp.376-380.

68. Helm, E., Lin, A.M., Baumgartner, D., Lin, A.C. and Küng, J. Towards the Use of Standardized Terms in Clinical Case Studies for Process Mining in Healthcare. *International Journal of Environmental Research and Public Health.* 2020, **17**(4), p.1348.

69. World Health Organization. *Cardiovascular Disease.* [Online]. 2017. [Accessed]. Available from: http://www.who.int/cardiovascular_diseases/en/

70. Kusuma, G.P., Hall, M., Gale, C.P. and Johnson, O.A. Process Mining in Cardiology: A Literature Review. *International Journal of Bioscience, Biochemistry and Bioinformatics.* 2018, **8**(4), pp.226-236.

71. Augusto, V., Xie, X., Prodel, M., Jouaneton, B. and Lamarsalle, L. Evaluation of discovered clinical pathways using process mining and joint agent-based discrete-event simulation. *Proceedings - Winter Simulation Conference.* 2017, (December), pp.2135-2146.

72. Fernandez-Llatas, C., Bayo, J.L., Martinez-Romero, A., Benedi, J.M., Traver, V., Bened´, J.M. and Traver, V. Interactive pattern recognition in cardiovascular disease management. A process mining approach. In: *2016/02//*: IEEE, pp.348-351.

73. Partington, A., Karnon, J., Wynn, M., Suriadi, S. and Ouyang, C. Process Mining for Clinical Processes: A Comparative Analysis of Four Australian Hospitals. *ACM Trans. Manag. Inform. Syst. Article.* 2015, **5**(19).

74. Fernandez-Llatas, C., Valdivieso, B., Traver, V. and Benedi, J.M. Using Process Mining for Automatic Support of Clinical Pathways Design. In: Fernández-Llatas, C. and García-Gómez, J. eds. New York: Humana Press, 2015, pp.79-88.

75. Dallagassa, M.R., dos Santos Garcia, C., Scalabrin, E.E., Ioshii, S.O. and Carvalho, D.R. Opportunities and challenges for applying process mining in healthcare: a systematic mapping study. *Journal of Ambient Intelligence and Humanized Computing.* 2021.

76. Homayounfar, P. Process mining challenges in hospital information systems. *Proceedings of the Federated Conference on Computer Science and Information Systems.* 2012, pp.1135-1140.

77. Dumas, M., La Rosa, M., Mendling, J. and Reijers, H.A. Introduction to Business Process Management. In: *Fundamentals of Business Process Management*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2013, pp.1-31.

78. Bose, R.P.J.C. and van der Aalst, W.M.P. Context Aware Trace Clustering: Towards Improving Process Mining Results. In: Philadelphia, PA: Society for Industrial and Applied Mathematics, 2009, pp.401-412.

79. Reijers, H.A., Mendling, J. and Dijkman, R.M. Human and automatic modularizations of process models to enhance their comprehension. *Information Systems.* 2011, **36**(5), pp.881-897.

80. Fox, F., Aggarwal, V.R., Whelton, H. and Johnson, O. A Data Quality Framework for Process Mining of Electronic Health Record Data. In: *2018 IEEE International Conference on Healthcare Informatics (ICHI)*, *4-7 June 2018*, 2018, pp.12-21.

81. Kurniati, A.P., McInerney, C., Zucker, K., Hall, G., Hogg, D. and Johnson, O. A Multi-level Approach for Identifying Process Change in Cancer Pathways. In: *Cham*. Springer International Publishing, 2019, pp.595-607.

82. Weiskopf, N.G. and Weng, C. Methods and dimensions of electronic health record data quality assessment: Enabling reuse for clinical research. *Journal of the American Medical Informatics Association.* 2013, **20**(1), pp.144-151.

83. NHS Health Research Authority. *Guidance for using patient data.* [Online]. 2021. [Accessed 10-Jul-2021]. Available from: https://www.hra.nhs.uk/

84. Rafiei, M., von Waldthausen, L. and van der Aalst, W.M.P. Supporting Confidentiality in Process Mining Using Abstraction and Encryption. In: *Cham*. Springer International Publishing, 2020, pp.101-123.

85. Rafiei, M. and van der Aalst, W.M.P. Privacy-Preserving Data Publishing in Process Mining. In: *Cham*. Springer International Publishing, 2020, pp.122-138.

86. Kusuma, G., Kurniati, A., McInerney, C.D., Hall, M., Gale, C.P. and Johnson, O. Process Mining of Disease Trajectories in MIMIC-III: A Case Study. In: *LNBIP 2nd International Conference on Process Mining (ICPM 2020)*, *Virtual conference managed by the University of Padua*. Springer, Verlag, 2020.

87. de Toledo, P., Joppien, C., Sesmero, M.P. and Drews, P. Mining Disease Courses across Organizations: A Methodology Based on Process Mining of Diagnosis Events Datasets. In: *2019 41st Annu. Int. Conf. IEEE Eng. Med. Biol. Soc. (EMBC)*: IEEE, pp.354-357.

88. De Oliveira, H., Prodel, M., Lamarsalle, L., Inada-Kim, M., Ajayi, K., Wilkins, J., Sekelj, S., Beecroft, S., Snow, S., Slater, R. and Orlowski, A. "Bow-tie" optimal pathway discovery analysis of sepsis hospital admissions using the Hospital Episode Statistics database in England. *JAMIA Open.* 2020, **3**(3), pp.439-448.

89. Bellodi, E., Riguzzi, F. and Lamma, E. Statistical relational learning for workflow mining. *Intelligent Data Analysis.* 2016, **20**, pp.515-541.

90. van Dongen, B.F., Carmona, J. and Chatain, T. A Unified Approach for Measuring Precision and Generalization Based on Anti-alignments. In: *Cham*. Springer International Publishing, 2016, pp.39-56.

91. Pareto, V. Traité de Sociologie Générale. 1re édn. *Librairie Droz.* 1917.

92. von Rosing, M., Scheer, A.-W., Zachman, J.A., Jones, D.T., Womack, J.P. and von Scheel, H. Phase 3: Process Concept Evolution. In: von Rosing, M.Scheer, A.-W.von Scheel, H. eds. *The Complete Business Process Handbook*. Boston: Morgan Kaufmann, 2015, pp.37-77.

93. Center for Disease Control and Prevention. Section 5: Measures of Association. In: *Principles of Epidemiology in Public Health Practice, Third Edition*
An Introduction to Applied Epidemiology and Biostatistics. 2012.

94. Tenny S, H.M. *Relative Risk.* [Online]. 2020. [Accessed]. Available from: https://www.ncbi.nlm.nih.gov/books/NBK430824/

95. Turhan, N.S. Karl Pearsons chi-square tests. *Educational Research and Reviews.* 2020, **15**(9), pp.575-580.

96. Fisher, R.A. Statistical Methods for Research Workers. In: Kotz, S. and Johnson, N.L. eds. *Breakthroughs in Statistics: Methodology and Distribution*. New York, NY: Springer New York, 1992, pp.66-70.

97. IBM. *Binomial Test.* [Online]. 2021. [Accessed 15-June-2021]. Available from: https://www.ibm.com

98. Abdi, H. Binomial distribution: Binomial and sign tests. *Encyclopedia of measurement and statistics.* 2007, **1**.

99. Hastie, T., Tibshirani, R., Friedman, J.H. and Friedman, J.H. *The elements of statistical learning: data mining, inference, and prediction.* Springer, 2009.

100. Hart, P.E., Stork, D.G. and Duda, R.O. *Pattern classification.* Wiley Hoboken, 2000.

101. Berrar, D. Cross-Validation. In: *Encyclopedia of Bioinformatics and Computational Biology*. Elsevier, 2018.

102. Simon, R. Supervised analysis when the number of candidate features (p) greatly exceeds the number of cases (n). *ACM SIGKDD Explorations Newsletter.* 2003, **5**(2), pp.31-36.

103. Picard, R.R. and Cook, R.D. Cross-Validation of Regression Models. *Journal of the American Statistical Association.* 1984, **79**(387), pp.575-583.

104. Kuhn, M. and Johnson, K. *Applied predictive modeling.* Springer New York, 2013.

105. Bishop, C.M. and Nasrabadi, N.M. *Pattern recognition and machine learning.* Springer New York, 2006.

106. Health, C.f.M.a.M.S.C.a.t.N.C.f. and Statistics (NCHS). *ICD-9-CM Official Guidelines for Coding and Reporting.* [Online]. 2011. [Accessed]. Available from: https://www.cdc.gov/nchs/data/icd/icd9cm_guidelines_2011.pdf

107. World Health Organisation. *International statistical classification of diseases and related health problems - 10th revision.* 2011.

108. Dunford, R., Su, Q. and Tamang, E. The pareto principle. *The Plymouth Student Scientist.* 2014, **7**(1), pp.140-148.

109. Müller, F., Dormann, H., Pfistermeister, B., Sonst, A., Patapovas, A., Vogler, R., Hartmann, N., Plank-Kiegele, B., Kirchner, M., Bürkle, T. and Maas, R. Application of the Pareto principle to identify and address drug-therapy safety issues. *European Journal of Clinical Pharmacology.* 2014, **70**(6), pp.727-736.

110. Morris, J.A. and Gardner, M.J. Calculating Confidence Intervals For Relative Risks (Odds Ratios) And Standardised Ratios And Rates. *British Medical Journal (Clinical Research Edition).* 1988, **296**(6632), pp.1313-1316.

111. Virtanen, P., Gommers, R., Oliphant, T.E., Haberland, M., Reddy, T., Cournapeau, D., Burovski, E., Peterson, P., Weckesser, W., Bright, J., van der Walt, S.J., Brett, M., Wilson, J., Millman, K.J., Mayorov, N., Nelson, A.R.J., Jones, E., Kern, R., Larson, E., Carey, C.J., Polat, İ., Feng, Y., Moore, E.W., VanderPlas, J., Laxalde, D., Perktold, J., Cimrman, R., Henriksen, I., Quintero, E.A., Harris, C.R., Archibald, A.M., Ribeiro, A.H., Pedregosa, F., van Mulbregt, P., Vijaykumar, A., Bardelli, A.P., Rothberg, A., Hilboll, A., Kloeckner, A., Scopatz, A., Lee, A., Rokem, A., Woods, C.N., Fulton, C., Masson, C., Häggström, C., Fitzgerald, C., Nicholson, D.A., Hagen, D.R., Pasechnik, D.V., Olivetti, E., Martin, E., Wieser, E., Silva, F., Lenders, F., Wilhelm, F., Young, G., Price, G.A., Ingold, G.-L., Allen, G.E., Lee, G.R., Audren, H., Probst, I., Dietrich, J.P., Silterra, J., Webber, J.T., Slavič, J., Nothman, J., Buchner, J., Kulick, J., Schönberger, J.L., de Miranda Cardoso, J.V., Reimer, J., Harrington, J., Rodríguez, J.L.C., Nunez-Iglesias, J., Kuczynski, J., Tritz, K., Thoma, M., Newville, M., Kümmerer, M., Bolingbroke, M., Tartre, M., Pak, M., Smith, N.J., Nowaczyk, N., Shebanov, N., Pavlyk, O., Brodtkorb, P.A., Lee, P., McGibbon, R.T., Feldbauer, R., Lewis, S., Tygier, S., Sievert, S., Vigna, S., Peterson, S., More, S., Pudlik, T., Oshima, T., Pingel, T.J., Robitaille, T.P., Spura, T., Jones, T.R., Cera, T., Leslie, T., Zito, T., Krauss, T., Upadhyay, U., Halchenko, Y.O., Vázquez-Baeza, Y. and SciPy, C. SciPy 1.0: fundamental algorithms for scientific computing in Python. *Nature Methods.* 2020, **17**(3), pp.261-272.

112. Kang, S.-H. and Ahn, C.W. Tests for the homogeneity of two binomial proportions in extremely unbalanced 2 x 2 contingency tables. *Statistics in Medicine.* 2008, **27**(14), pp.2524-2535.

113. Bender, R. and Lange, S. Adjusting for multiple testing—when and how? *Journal of Clinical Epidemiology.* 2001, **54**(4), pp.343-349.

114. Rothman, K.J. No Adjustments Are Needed for Multiple Comparisons. *Epidemiology.* 1990, **1**(1), pp.43-46.

115. Mannhardt, F., De Leoni, M. and Reijers, H.A. Heuristic mining revamped: An interactive, data-Aware, and conformance-Aware miner. In: *BPM 2017*: CEUR-WS.org, pp.1-5.

116. Fluxicon, B.V. *Disco.* [Online]. 2019. [Accessed]. Available from: https://fluxicon.com/disco/

117. Adriansyah, A. *Replay a Log on Petri Net for Conformance Analysis-plugin.* 2012.

118. Adriansyah, A., Dongen, B.F.v. and Aalst, W.M.P.v.d. Conformance Checking Using Cost-Based Fitness Analysis. In: *2011 IEEE 15th International Enterprise Distributed Object Computing Conference*, *29 Aug.-2 Sept. 2011*, 2011, pp.55-64.

119. Adriansyah, A., Munoz-Gama, J., Carmona, J., van Dongen, B.F. and van der Aalst, W.M.P. Measuring precision of modeled behavior. *Information Systems and e-Business Management.* 2015, **13**(1), pp.37-67.

120. Kusuma, G., Sykes, S., McInerney, C. and Johnson, O. *Resource of Process Mining for Disease Trajectory Mining*, 2019.

121. Cordella, L.P., Foggia, P., Sansone, C. and Vento, M. An improved algorithm for matching large graphs. In: *2001*, pp.149-159.

122. Siggaard, T., Reguant, R., Jørgensen, I.F., Haue, A.D., Lademann, M., Aguayo-Orozco, A., Hjaltelin, J.X., Jensen, A.B., Banasik, K. and Brunak, S. Disease trajectory browser for exploring temporal, population-wide disease progression patterns in 7.2 million Danish patients. *Nature Communications.* 2020, **11**(1), p.4952.

123. Beck, M.K., Boeck Jensen, A., Bach Nielsen, A., Perner, A., Moseley, L. and Brunak, S. Diagnosis trajectories of prior multi-morbidity predict sepsis mortality OPEN. *Nature Publishing Group.* 2016.

124. Process-Oriented Data Science for Healthcare. *PODS4H.* [Online]. 2019. [Accessed]. Available from: https://www.pods4h.com

125. Goldberger, A.L., Amaral, L.A.N., Glass, L., Hausdorff, J.M., Ivanov, P.C., Mark, R.G., Mietus, J.E., Moody, G.B., Peng, C.-K. and Stanley, H.E. PhysioBank, PhysioToolkit, and PhysioNet. *Circulation.* 2000, **101**(23), pp.e215-e220.

126. Herbert, A., Wijlaars, L., Zylbersztejn, A., Cromwell, D. and Hardelid, P. Data Resource Profile: Hospital Episode Statistics Admitted Patient Care (HES APC). *International journal of epidemiology.* 2017, **46**(4), pp.1093-1093i.

127. Health and Social Care Information Centre. *The HES Processing Cycle and Data Quality*. 2014. 6 June 2019. Available from:

http://content.digital.nhs.uk/media/1366/The-HES-processing-cycle-and-HES-data-quality/pdf/The_HES_Processing_Cycle_and_HES_Data_Quality_v3.pdf

128. Health and Social Care Information Centre. *Cleaning Rules: Admitted Patient Care*. 2016. Available from: http://content.digital.nhs.uk/media/1367/HES-Hospital-Episode-Statistics-Inpatient-cleaning-rules/pdf/HES__APC_052015b.pdf

129. Health and Social Care Information Centre. *Data Quality Checks Performed on SUS and HES Data*. 2014. Available from: http://content.digital.nhs.uk/media/13655/Data-quality-checks-performed-on-SUS-and-HES-data/pdf/Data_quality_checks_performed_on_SUS_and_HES_data_v2.pdf

130. Deeks, A., Lombard, C., Michelmore, J. and Teede, H. The effects of gender and age on health related behaviors. *BMC Public Health.* 2009, **9**(1), p.213.

131. World Health Organisation. *Gender and Health.* [Online]. 2021. [Accessed 3 September]. Available from: https://who.int

132. Mannhardt, F. Tools & Software — ProM — Event Log Explorer. 2018. [Online]. Available from: https://fmannhardt.de/blog/software/prom/explorer

133. Asaria, P., Elliott, P., Douglass, M., Obermeyer, Z., Soljak, M., Majeed, A. and Ezzati, M. Acute myocardial infarction hospital admissions and deaths in England: a national follow-back and follow-forward record-linkage study. *The Lancet Public Health.* 2017, **2**(4), pp.e191-e201.

134. Hall, M., Dondo, T.B., Yan, A.T., Mamas, M.A., Timmis, A.D., Deanfield, J.E., Jernberg, T., Hemingway, H., Fox, K.A.A. and Gale, C.P. Multimorbidity and survival for patients with acute myocardial infarction in England and Wales: Latent class analysis of a nationwide population-based cohort. *PLoS Medicine.* 2018, **15**(3).

135. Sakr, Y., Jaschinski, U., Wittebole, X., Szakmany, T., Lipman, J., Ñamendys-Silva, S.A., Martin-Loeches, I., Leone, M., Lupu, M.-N., Vincent, J.-L. and Icon Investigators, I. Sepsis in Intensive Care Unit Patients: Worldwide Data From the Intensive Care over Nations Audit. *Open forum infectious diseases.* 2018, **5**(12), pp.ofy313-ofy313.

136. Millett, E.R.C., Peters, S.A.E. and Woodward, M. Sex differences in risk factors for myocardial infarction: cohort study of UK Biobank participants. *BMJ.* 2018, **363**, p.k4247.

137. Wáng, Y.X., He, J., Zhang, L., Li, Y., Zhao, L., Liu, H., Yang, L., Zeng, X.J., Yang, J., Peng, G.M., Ahuja, A. and Yang, Z.H. A higher aneurysmal subarachnoid hemorrhage incidence in women prior to menopause: a retrospective analysis of 4,895 cases from eight hospitals in China. *Quant Imaging Med Surg.* 2016, **6**(2), pp.151-156.

138. Mehta, R.L., Bouchard, J., Soroko, S.B., Ikizler, T.A., Paganini, E.P., Chertow, G.M., Himmelfarb, J. and Program to Improve Care in Acute Renal Disease Study, G. Sepsis as a cause and consequence of acute kidney injury:

Program to Improve Care in Acute Renal Disease. *Intensive care medicine.* 2011, **37**(2), pp.241-248.

139. Peerapornratana, S., Manrique-Caballero, C.L., Gómez, H. and Kellum, J.A. Acute kidney injury from sepsis: current concepts, epidemiology, pathophysiology, prevention and treatment. *Kidney international.* 2019, **96**(5), pp.1083-1099.

140. Eden, S.V., Meurer, W.J., Sánchez, B.N., Lisabeth, L.D., Smith, M.A., Brown, D.L. and Morgenstern, L.B. Gender and ethnic differences in subarachnoid hemorrhage. *Neurology.* 2008, **71**(10), pp.731-735.

141. Parris, G. Towards a coordinated approach for management information in the NHS. *Health Libraries Review.* 1986, **3**(2), pp.82-93.

142. Boyd, A., Cornish, R., Johnson, L., Simmonds, S., Syddall, H., Westbury, L., Cooper, C. and Macleod, J. *Understanding Hospital Episode Statistics (HES) Resource report.* 2018.

143. Black, D. Data for management: the Körner Report. 1982, (0267-0623 (Print)).

144. NHS Digital. *Hospital Episode Statistics (HES).* [Online]. 2021. [Accessed]. Available from: https://digital.nhs.uk/data-and-information/data-tools-and-services/data-services/hospital-episode-statistics

145. Sinha, S., Peach, G., Poloniecki, J.D., Thompson, M.M. and Holt, P.J. Studies using English administrative data (Hospital Episode Statistics) to assess health-care outcomes—systematic review and recommendations for reporting. *European Journal of Public Health.* 2012, **23**(1), pp.86-92.

146. Herrett, E., Gallagher, A.M., Bhaskaran, K., Forbes, H., Mathur, R., van Staa, T. and Smeeth, L. Data Resource Profile: Clinical Practice Research Datalink (CPRD). *International journal of epidemiology.* 2015, **44**(3), pp.827-836.

147. NHS Digital. *Spine.* [Online]. 2021. [Accessed]. Available from: https://digital.nhs.uk/services/spine

148. NHS Digital. *HES Autocleanse Dictionary - Accident and Emergency, Admitted Patient Care and Outpatient Care.* 2017.

149. Partridge, N. *Review of data releases by the NHS Information Centre.* www.gov.uk, 2014.

150. NHS Digital. *Hospital Episode Statistics data changes in 2021.* [Online]. 2021. [Accessed]. Available from: https://digital.nhs.uk/data-and-information/data-tools-and-services/data-services/hospital-episode-statistics/hospital-episode-statistics-data-changes-in-2021

151. HSCIC. Methodology to create provider and CIP spells from HES APC data. [Online]. 2014. Available from: webarchie.nationalarchive.gov.uk

152. Fox, F., Aggarwal, V.R., Whelton, H. and Johnson, O. A data quality framework for process mining of electronic health record data. In: *2018*: IEEE International Conference, 2018, pp.12-21.

153. Mans, R.S., van der Aalst, W.M.P. and Vanwersch, R.J.B. *Process Mining in Healthcare Evaluating and Exploiting Operational Healthcare Processes.* 1 ed. Springer International Publishing, 2015.

154. Bose, R.P.J.C., Mans, R.S. and Van Der Aalst, W.M.P. Wanna Improve Process Mining Results? It's High Time We Consider Data Quality Issues Seriously. *Proceedings of the 2013 IEEE Symposium on Computational Intelligence and Data Mining, CIDM 2013 - 2013 IEEE Symposium Series on Computational Intelligence, SSCI 2013.* 2013, (1), pp.127-134.

155. Vandenbroucke, J.P., von Elm, E., Altman, D.G., Gøtzsche, P.C., Mulrow, C.D., Pocock, S.J., Poole, C., Schlesselman, J.J., Egger, M. and for the, S.I. Strengthening the Reporting of Observational Studies in Epidemiology (STROBE): Explanation and Elaboration. *PLoS Medicine.* 2007, **4**(10), p.e297.

156. NHS Digital. *Hospital Episode Statistics Data Dictionary.* NHS Digital, 2017.

157. NHS Digital. *Methodology for identifying and removing duplicate records from the HES dataset.* NHS Digital, 2016.

158. Alrawashdeh, H.M., Naser, A.Y., Alwafi, H., AbuAlhommos, A.K., Jalal, Z., Paudyal, V., Abdulmannan, D.M., Hassanin, F.F., Hemmo, S.I. and Al Sarireh, F. Trends in Hospital Admission Due to Diseases of the Eye and Adnexa in the Past Two Decades in England and Wales: An Ecological Study. *Int J Gen Med.* 2022, **15**, pp.1097-1110.

159. Teringova, E. and Tousek, P. Apoptosis in ischemic heart disease. *Journal of Translational Medicine.* 2017, **15**(1), p.87.

160. Restrepo, M.I., Sibila, O. and Anzueto, A. Pneumonia in Patients with Chronic Obstructive Pulmonary Disease. *Tuberculosis and respiratory diseases.* 2018, **81**(3), pp.187-197.

161. Hartley, B.F., Barnes, N.C., Lettis, S., Compton, C.H., Papi, A. and Jones, P. Risk factors for exacerbations and pneumonia in patients with chronic obstructive pulmonary disease: a pooled analysis. *Respiratory Research.* 2020, **21**(1), p.5.

162. Kurniati, A.P., Rojas, E., Hogg, D. and Johnson, O. The assessment of data quality issues for process mining in healthcare using MIMIC-III , a publicly available e-health record database. *Health informatics journal.* 2017, **25**(4), pp.1878-1893.

163. Hemingway, H., Feder, G., Fitzpatrick, N., Denaxas, S., Shah, A. and Timmis, A. Using nationwide ‘big data’ from linked electronic health records to help improve outcomes in cardiovascular diseases: 33 studies using methods from epidemiology, informatics, economics and social science in the ClinicAl disease research using LInked Bespoke studies and Electronic health Records

(CALIBER) programme. *Programme Grants for Applied Research.* 2017, **5**, pp.1-330.

164. Raghupathi, W. and Raghupathi, V. Big data analytics in healthcare: promise and potential. *Health Information Science and Systems.* 2014, **2**(1), p.3.

165. Health, C.f.M.a.M.S.C.a.t.N.C.f. and Statistics (NCHS). *ICD-10-CM Official Guidelines for Coding and Reporting.* 2021.

166. Berry, L.L., Deming, K.A. and Danaher, T.S. Improving Nonclinical and Clinical-Support Services: Lessons From Oncology. *Mayo Clin Proc Innov Qual Outcomes.* 2018, **2**(3), pp.207-217.

List of Abbreviations

AKI	:	Acute Kidney Injury
AMI	:	Acute Myocardial Infarction
ANST	:	Any non-significant traces
APC	:	Admitted Patient Care
BC	:	Before Christ
BHF	:	British Heart Foundation
BIDMC	:	Beth Israel Deaconess Medical Center
BPA-H	:	Business Process Analysis in Healthcare
CCG	:	Clinical Commissioning Group
CDS	:	Commissioning Data Set
CI	:	Confidence Interval
CM	:	Clinical Modification
COPD	:	Chronic obstructive pulmonary disease
COVID	:	Coronavirus disease
CP-DQF	:	Care Pathway Data Quality Framework
CPRD	:	Clinical Practice Research Datalink
CVD	:	Cardiovascular disease
CVEPI	:	Cardiovascular Episodes
DARS	:	Data Access Request Service
DF	:	Degree of freedom
DFG	:	Directly-follows graph
iDHM	:	interactive Data-aware Heuristics Miner
DISDATE	:	Discharge date
DISDEST	:	Discharge destination
DISMETH	:	Discharge method
DOB	:	Date of birth
DQF	:	Data Quality Framework
DSA	:	Data Sharing Agreement
DSFC	:	Data Sharing Framework Contract
EHR	:	Electronic Health Records
EPIEND	:	Episode end
EPIKEY	:	Episode key
EPIORDER	:	Episode order
EPISTART	:	Episode start
EPISTAT	:	Episode status
EPITYPE	:	Episode type
EPS	:	Electronic Prescriptions Service
ETL	:	Extract Transform Load
FAE	:	Finished Admission Episodes
FCE	:	Finished Consultant Episode
GB	:	Giga byte
GLOW	:	Global Learning Week

GP	:	General Practitioner
HEALTHINF	:	Heatlth Informatics
HES	:	Hospital Episode Statistics
HESID	:	HES Identifier
HIPAA	:	Health Insurance Portability and Accountability Act
HOMEADD	:	Home address
HSCIC	:	Health & Social Care Information Care
ICD	:	International Classification of Disease
ICPM	:	International Conference of Process Mining
IJBBB	:	International Journal of Bioscience, Biochemistry, and Bioinformatics
IM	:	Inductive Miner
IQR	:	Interquartile range
IRC	:	Integrated Research Campus
LCI	:	Lower confidence interval
LHS	:	Learning health system
LIDA	:	Leeds Instituite of Data Analytics
MIE	:	Medical Informatics Europe
MIMIC	:	Medical Information Mart for Intensive Care
MIT	:	Massachusetts Institute of Technology
MPS	:	Master Person Service
MXML	:	Macromedia FleX Markup Language
MYADMIDATE	:	Month Year Admission Date
MYDOB	:	Month Year Date of birth
MYEPIEND	:	Month Year Episode End
MYEPISTART	:	Month Year Episode Start
NCHS	:	National Center for Health Statistics
NHS	:	National Health Service
ODBC	:	Open Database Connectivity
OR	:	Odd Ratio
OS	:	Operating System
PDM	:	Process Diagnostic Method
PM^2	:	Process Mining Project Methodology
PODS	:	Process-Oriented Data Science
RQ	:	Research Question
RR	:	Relative Risk
SD	:	Standard Deviation
SEED	:	Secure Electronic Environment for Data
SEQ	:	Sequence
SNOMED-CT	:	Systematized Nomenclature of Medicine – Clinical Term
SPELBGIN	:	Spell begin
SPELEND	:	Spell end
SQL	:	Structured Query Language

STROBE	:	Strengthening the Reporting of Observational Studies in Epidemiology
SUS	:	Secondary User Service
UK	:	United Kingdom
US	:	United States
USA	:	United States of America
WEKA	:	Waikato Environment for Knowledge Analysis
WHO	:	World Health Organization
XES	:	eXtensible Event Stream
XML	:	eXtended Markup Language

Appendix A
NHS England HES-APC Data Dictionary

Dataset	Category	Field	Field name	NHS Field Name	Format	Availability	Description	Value
APC	Admissions; Period of Care	ADMIMETH	Method of admission	Admission Method (Hospital Provider Spell) (V6-1) Admission Method Code (Hospital Provider Spell) (V6-2)	2n	1989-90 onwards	This field contains a code which identifies how the patient was admitted to hospital. Admimeth is recorded on the first and also all subsequent episodes within the spell (ie where the spell is made up of more than one episode).	Elective Admission, when the decision to admit could be separated in time from the actual admission: 11 = Waiting list. . A Patient admitted electively from a waiting list having been given no date of admission at a time a decision was made to admit 12 = Booked. A Patient admitted having been given a date at the time the decision to admit was made, determined mainly on the grounds of resource availability 13 = Planned. A Patient admitted, having been given a date or approximate date at the time that the decision to admit was made. This is usually part of a planned sequence of clinical care determined mainly on social or clinical criteria (e.g. check cystoscopy)". A planned admission is one where the date of admission is determined by the needs of the treatment, rather than by the availability of resources. Note that this does not include a transfer from another Hospital Provider (see 81 below). Emergency Admission, when admission is unpredictable and at short notice because of clinical need: 21 = Accident and emergency or dental casualty department of the Health Care Provider 22 = General Practitioner: after a request for immediate admission has been made direct to a Hospital Provider, i.e. not through a Bed bureau, by a General Practitioner: or deputy 23 = Bed bureau 24 = Consultant Clinic, of this or another Health Care Provider 25 = Admission via Mental Health Crisis Resolution Team (available from 2013/14) 2A = Accident and Emergency Department of another provider where the patient had not been admitted (available from 2013/14) 2B = Transfer of an admitted patient from another Hospital Provider in an emergency (available from 2013/14) 2C = Baby born at home as intended (available from 2013/14) 2D = Other emergency admission (available from 2013/14) 28 = Other means, examples are: - Admitted from the Accident and Emergency Department of another provider where they had not been admitted - Transfer of an admitted patient from another Hospital Provider in an emergency
APC	Discharges; Period of Care	DISMETH	Method of discharge	Discharge Method (Hospital Provider Spell) (V6-1) Discharge Method Code (Hospital Provider Spell) (V6-2)	1n	1989-90 onwards	This field contains a code which defines the circumstances under which a patient left hospital. For the majority of patients this is when they are discharged by the consultant. This field is only completed for the last episode in a spell.	1 = Discharged on clinical advice or with clinical consent 2 = Self discharged, or discharged by a relative or advocate 3 = Discharged by a mental health review tribunal, the Home Secretary or a court 4 = Died 5 = Baby was still born 8 = Not applicable: patient still in hospital 9 = Not known: a validation error

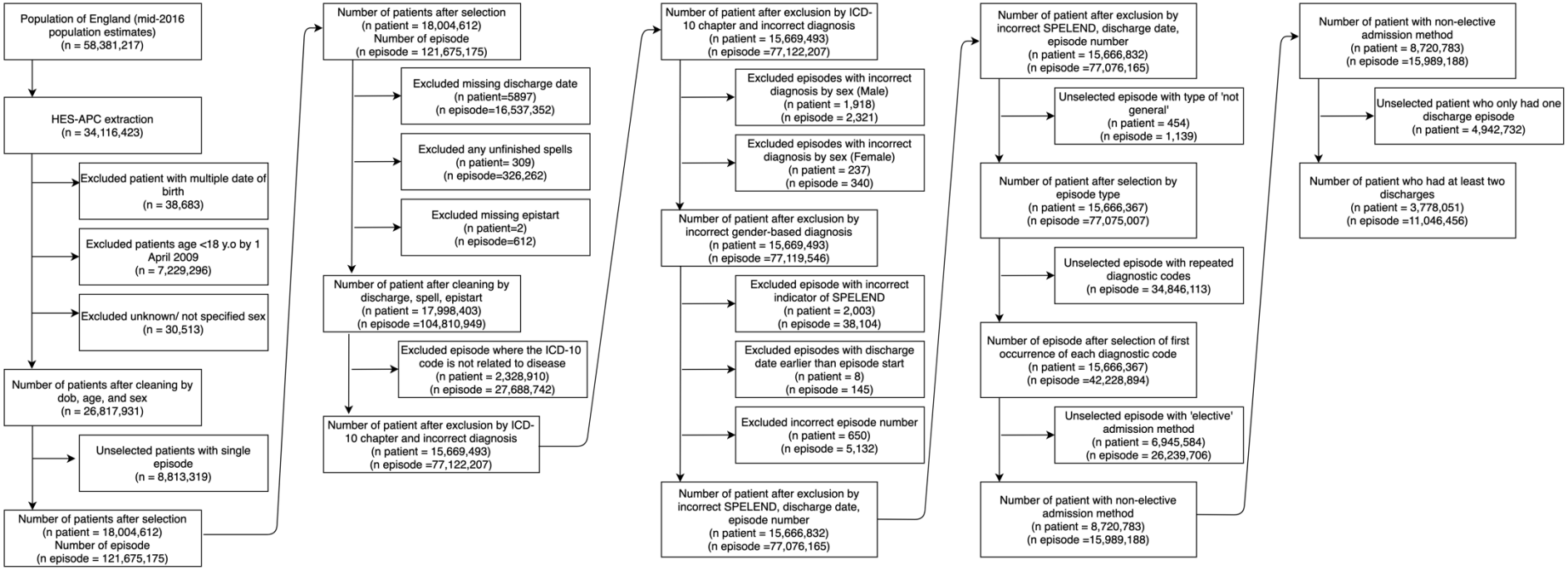
Dataset	Category	Field	Field name	NHS Field Name	Format	Availability	Description	Value
APC	Clinical	DIAG_NN	All Diagnosis codes	Primary Diagnosis (ICD) Secondary Diagnosis (ICD)	6an	1989-90 onwards	There are twenty fields (fourteen before April 2007 and seven before April 2002), diag_01 to diag_20, which contain information about a patient's illness or condition. The field diag_01 contains the primary diagnosis. The other fields contain secondary/subsidiary diagnoses. The codes are defined in the International Statistical Classification of Diseases, Injuries and Causes of Death. HES records currently use the tenth revision (ICD-10). Prior to April 1995, the ninth revision was used (ICD-9). Diagnosis codes start with a letter and are followed by two or three digits. The third digit identifies variations on a main diagnosis code containing two digits. The third digit is preceded by a full stop in ICD-10, but this is not stored in the field.	annnna = A valid ICD-9 or ICD-10 diagnosis code annnnn = A valid ICD-9 or ICD-10 diagnosis code Null = Not applicable R96X - Not known R69X6 - Null (Primary diagnosis) R69X8 - Invalid R69X3 = Invalid (External Cause code entered as Primary Diagnosis)
APC	Episodes and spells; Period of care	EPIEND	Date episode ended	End Date (Episode)	dd/mm/yy yy (Date)	1989-90 onwards	This field contains the date on which a patient left the care of a particular consultant, for one of the following reasons: Patient discharged from hospital (includes transfers) or moved to the care of another consultant. A null entry either indicates that the episode was unfinished at the end of the data year, or the date was unknown.	2012/13 onwards: 01/01/1800 - Null date submitted 01/01/1801 - Invalid date submitted 1989/90 to 2011/12: 01/01/1600 – Null date submitted 15/10/1582 – Invalid date submitted
APC	Episodes and spells; Period of care	EPISTART	Date episode started	Start Date (Episode)	dd/mm/yy yy (Date)	1989-90 onwards	This field contains the date on which a patient was under the care of a particular consultant. If a patient has more than one episode in a spell, for each new episode there is a new value of epistart. However, the admission date which is copied to each new episode in a spell will remain unchanged and will be equal to the episode start date of the first episode in hospital.	2012/13 onwards: 01/01/1800 - Null date submitted 01/01/1801 - Invalid date submitted 1989/90 to 2011/12: 01/01/1600 – Null date submitted 15/10/1582 – Invalid date submitted
APC	Admissions; Period of Care	ADMIDATE	Date of admission	Start Date (Hospital Provider Spell)	dd/mm/yy yy (Date)	1989-90 onwards	This field contains the date the patient was admitted to hospital at the start of a hospital spell. Admidate is recorded on all episodes within a spell.	2012/13 onwards: 01/01/1800 - Null date submitted 01/01/1801 - Invalid date submitted 1989/90 to 2011/12: 01/01/1600 – Null date submitted 15/10/1582 – Invalid date submitted
APC	Patient Data	DOB	Date of birth - patient	Person Birth Date	dd/mm/yy yy (Date)	1989-90 onwards	This field contains the patient's date of birth. For most enquiries the field startage (age at start of episode) is used. The Date or birth - patient (dob) field contains sensitive data. Access to it requires the approval of the Confidentiality Advisory Group (CAG).	2012/13 onwards: 01/01/1800 - Null date submitted 01/01/1801 - Invalid date submitted 1989/90 to 2011/12: 01/01/1600 – Null date submitted 15/10/1582 – Invalid date submitted

Continued…

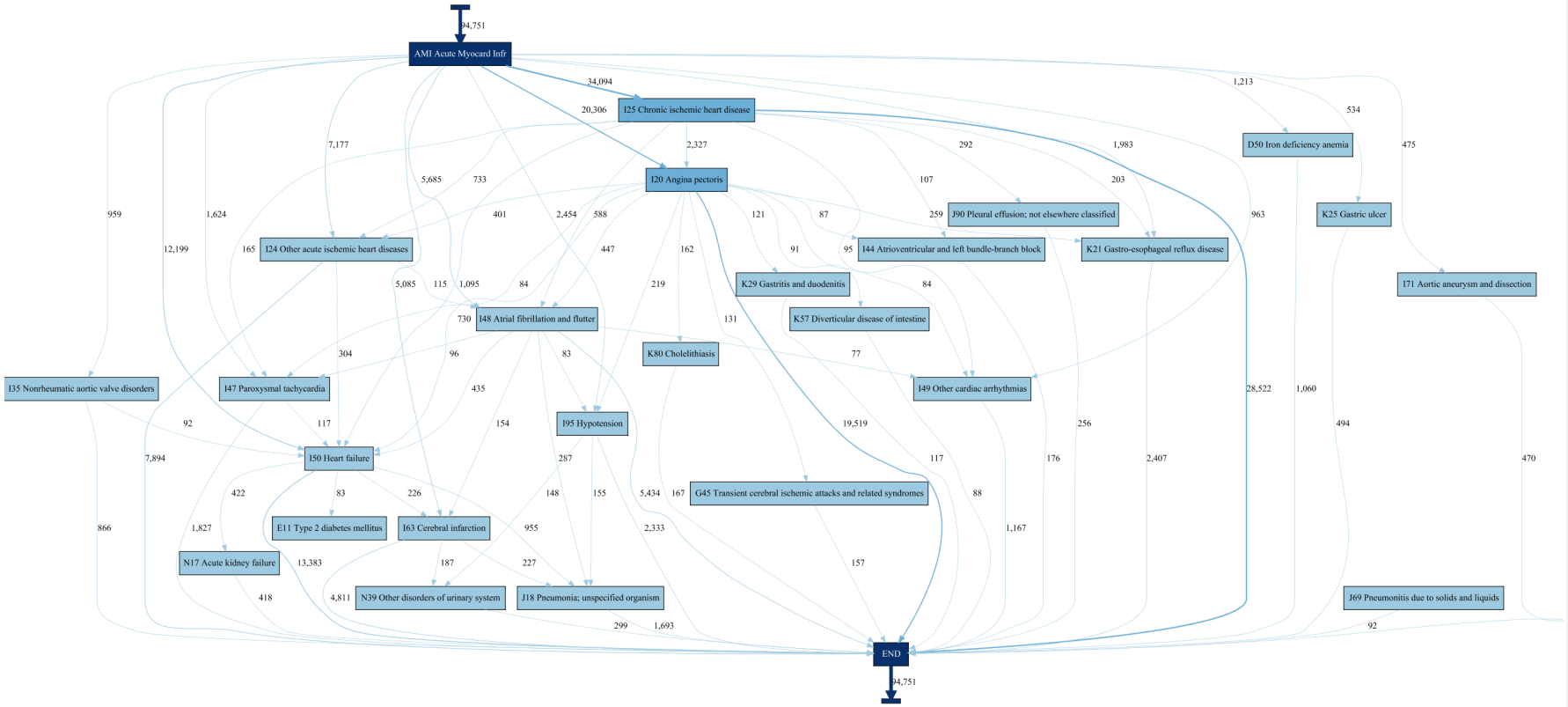
Dataset	Category	Field	Field name	NHS Field Name	Format	Availability	Description	Value
APC	Discharges; Period of Care	DISDATE	Date of discharge	Discharge Date (Hospital Provider Spell)	dd/mm/yy yy (Date)	1989-90 onwards	This field contains the date on which the patient was discharged from hospital. ; It is only present in the record for the last episode of a spell.	2012/13 onwards: 01/01/1800 - Null date submitted 01/01/1801 - Invalid date submitted 1989/90 to 2011/12: 01/01/1600 – Null date submitted 15/10/1582 – Invalid date submitted
APC	Episodes and spells; Period of care	EPIORDER	Episode order	Episode Number	2n	1989-90 onwards	This field contains the number of the episode within the current spell. All spells start with an episode where epiorder is 01. Many spells finish with this episode, but if the patient moves to the care of another consultant, a new episode begins. Episode numbers increase by 1 for each new episode until the patient is discharged (this includes transfers to another NHS trust or primary care trust - ie the first episode in the new trust will have epiorder 01). If the same patient returns for a different spell in hospital, epiorder is again set to 01. Admissions are calculated by counting the number of times epiorder is 01. When studying long term care, remember that it is not unusual to transfer psychiatric patients from one hospital to another.	2n = The number of the episode in the sequence of episodes from 01-87 98 = Not applicable 99 = Not known: a validation error Null = Not applicable: other maternity event
APC	Episodes and spells; Period of care	EPISTAT	Episode status	N/A	1n	1989-90 onwards	This field tells you whether the episode had finished before the end of the HES data-year (ie whether the episode was still 'live' at midnight on 31 March). For example, if a patient was admitted on 25 March 2005 and was not discharged (or transferred to the care of another consultant) until 4 April 2005, there will be a record describing the unfinished episode (episode status = 1) in the 2004-05 data, and a separate record describing the finished episode (episode status = 3) in the 2005-06 data. Because hospital providers are advised not to include clinical data (diagnosis and operation codes) in unfinished records, these are normally excluded from analyses. Also, if unfinished episodes are included in time series analyses - where data for more than one year is involved - there is a danger of counting the same episode twice.	1 = Unfinished 3 = Finished 9 = Derived unfinished (not present on processed data)

Dataset	Category	Field	Field name	NHS Field Name	Format	Availability	Description	Value
APC	Patient Data	PSEUDO_HESID	Pseudonymised HES ID	N/A	32an	2007-08 onwards	This field contains a unique identifier for each individual patient. This allows an individual's care to be tracked across years and continuous periods to be identified. This is a pseudonymised version of the HES ID field based on an updated matching algorithm, which supersedes and is compatible with the original HES ID, which is no longer available.	32an = Pseudonymised HESID
APC	Patient Data	SEX	Sex of patient	Person Gender Current (V6-1) Person Gender Code Current (V6-2)	1n	1989-90 onwards	Defines the sex of the patient. The classification is phenotypical rather than genotypical, i.e. it does not provide codes for medical or scientific purposes. Notes: • National Code 'Not Known' means that the sex of a person has not been recorded • National Code 'Not Specified' means indeterminate, i.e. unable to be classified as either male or female.	From 1996-97 onwards: 1 = Male 2 = Female 9 = Not specified 0 = Not known Prior to April 1996: 1 = Male 2 = Female 3 = Indeterminate, including those undergoing sex change operations

Appendix B
STROBE diagram of data exclusion and selection of HES-APC data set



Appendix C
Directly-follows graph of disease trajectory model using the optimum directly-follow frequency



Appendix D
Post-AMI Disease Trajectories

The exceptional trajectories post-AMI.

Variant index	Exceptional trajectories
286	AMI Acute Myocard Infr->I48 Atrial fibrillation and flutter->J90 Pleural effusion; not elsewhere classified->J22 Unspecified acute lower respiratory infection
287	AMI Acute Myocard Infr->I48 Atrial fibrillation and flutter->I95 Hypotension->M25 Other joint disorder; not elsewhere classified
288	AMI Acute Myocard Infr->I48 Atrial fibrillation and flutter->J18 Pneumonia; unspecified organism->C34 Malignant neoplasm of bronchus and lung
289	AMI Acute Myocard Infr->I48 Atrial fibrillation and flutter->I49 Other cardiac arrhythmias->I50 Heart failure
290	AMI Acute Myocard Infr->I48 Atrial fibrillation and flutter->I47 Paroxysmal tachycardia->I95 Hypotension
291	AMI Acute Myocard Infr->I48 Atrial fibrillation and flutter->I62 Other and unspecified nontraumatic intracranial hemorrhage->J18 Pneumonia; unspecified organism
292	AMI Acute Myocard Infr->I48 Atrial fibrillation and flutter->I63 Cerebral infarction->N39 Other disorders of urinary system->J18 Pneumonia; unspecified organism
293	AMI Acute Myocard Infr->I48 Atrial fibrillation and flutter->G45 Transient cerebral ischemic attacks and related syndromes->M25 Other joint disorder; not elsewhere classified
294	AMI Acute Myocard Infr->I48 Atrial fibrillation and flutter->M10 Gout->N17 Acute kidney failure
295	AMI Acute Myocard Infr->I49 Other cardiac arrhythmias->B34 Viral infection of unspecified site
296	AMI Acute Myocard Infr->I24 Other acute ischemic heart diseases->I48 Atrial fibrillation and flutter->I49 Other cardiac arrhythmias
297	AMI Acute Myocard Infr->D50 Iron deficiency anemia->I50 Heart failure->N17 Acute kidney failure
298	AMI Acute Myocard Infr->D50 Iron deficiency anemia->C78 Secondary malignant neoplasm of resp and digestive organs
299	AMI Acute Myocard Infr->D50 Iron deficiency anemia->N93 Other abnormal uterine and vaginal bleeding
300	AMI Acute Myocard Infr->I20 Angina pectoris->K21 Gastro-esophageal reflux disease->M79 Oth and unsp soft tissue disorders; not elsewhere classified
301	AMI Acute Myocard Infr->I20 Angina pectoris->K21 Gastro-esophageal reflux disease->K80 Cholelithiasis
302	AMI Acute Myocard Infr->I20 Angina pectoris->K22 Other diseases of esophagus->A09 Infectious gastroenteritis and colitis; unspecified
303	AMI Acute Myocard Infr->I20 Angina pectoris->I48 Atrial fibrillation and flutter->I71 Aortic aneurysm and dissection
304	AMI Acute Myocard Infr->I20 Angina pectoris->I48 Atrial fibrillation and flutter->I50 Heart failure->N17 Acute kidney failure
305	AMI Acute Myocard Infr->I20 Angina pectoris->I48 Atrial fibrillation and flutter->K55 Vascular disorders of intestine
306	AMI Acute Myocard Infr->I20 Angina pectoris->I48 Atrial fibrillation and flutter->I74 Arterial embolism and thrombosis
307	AMI Acute Myocard Infr->I20 Angina pectoris->I48 Atrial fibrillation and flutter->I44 Atrioventricular and left bundle-branch block
308	AMI Acute Myocard Infr->I20 Angina pectoris->G45 Transient cerebral ischemic attacks and related syndromes->F03 Unspecified dementia

309	AMI Acute Myocard Infr->I20 Angina pectoris->G45 Transient cerebral ischemic attacks and related syndromes->N39 Other disorders of urinary system->J18 Pneumonia; unspecified organism
310	AMI Acute Myocard Infr->I20 Angina pectoris->G45 Transient cerebral ischemic attacks and related syndromes->N39 Other disorders of urinary system
311	AMI Acute Myocard Infr->I20 Angina pectoris->G45 Transient cerebral ischemic attacks and related syndromes->I67 Other cerebrovascular diseases
312	AMI Acute Myocard Infr->I20 Angina pectoris->G45 Transient cerebral ischemic attacks and related syndromes->G43 Migraine
313	AMI Acute Myocard Infr->I20 Angina pectoris->K80 Cholelithiasis->M79 Oth and unsp soft tissue disorders; not elsewhere classified
314	AMI Acute Myocard Infr->I20 Angina pectoris->K80 Cholelithiasis->K56 Paralytic ileus and intestinal obstruction without hernia
315	AMI Acute Myocard Infr->I20 Angina pectoris->I50 Heart failure->K55 Vascular disorders of intestine
316	AMI Acute Myocard Infr->I20 Angina pectoris->I50 Heart failure->J96 Respiratory failure; not elsewhere classified
317	AMI Acute Myocard Infr->I20 Angina pectoris->I50 Heart failure->E11 Type 2 diabetes mellitus
318	AMI Acute Myocard Infr->I20 Angina pectoris->K29 Gastritis and duodenitis->K70 Alcoholic liver disease
319	AMI Acute Myocard Infr->I20 Angina pectoris->I24 Other acute ischemic heart diseases->I48 Atrial fibrillation and flutter->I50 Heart failure
320	AMI Acute Myocard Infr->I20 Angina pectoris->I95 Hypotension->N39 Other disorders of urinary system->N17 Acute kidney failure
321	AMI Acute Myocard Infr->I20 Angina pectoris->I95 Hypotension->F03 Unspecified dementia
322	AMI Acute Myocard Infr->I20 Angina pectoris->I95 Hypotension->K59 Other functional intestinal disorders
323	AMI Acute Myocard Infr->I20 Angina pectoris->I95 Hypotension->M25 Other joint disorder; not elsewhere classified
324	AMI Acute Myocard Infr->I20 Angina pectoris->B34 Viral infection of unspecified site->K59 Other functional intestinal disorders
325	AMI Acute Myocard Infr->I20 Angina pectoris->I44 Atrioventricular and left bundle-branch block->N17 Acute kidney failure
326	AMI Acute Myocard Infr->I20 Angina pectoris->I44 Atrioventricular and left bundle-branch block->I24 Other acute ischemic heart diseases
327	AMI Acute Myocard Infr->I20 Angina pectoris->I44 Atrioventricular and left bundle-branch block->M25 Other joint disorder; not elsewhere classified
328	AMI Acute Myocard Infr->I20 Angina pectoris->I35 Nonrheumatic aortic valve disorders->I50 Heart failure
329	AMI Acute Myocard Infr->I20 Angina pectoris->I35 Nonrheumatic aortic valve disorders->I25 Chronic ischemic heart disease->I50 Heart failure->J18 Pneumonia; unspecified organism
330	AMI Acute Myocard Infr->I20 Angina pectoris->K40 Inguinal hernia->K59 Other functional intestinal disorders
331	AMI Acute Myocard Infr->I25 Chronic ischemic heart disease->I50 Heart failure->N17 Acute kidney failure->J18 Pneumonia; unspecified organism
332	AMI Acute Myocard Infr->I25 Chronic ischemic heart disease->I50 Heart failure->I63 Cerebral infarction->N39 Other disorders of urinary system
333	AMI Acute Myocard Infr->I25 Chronic ischemic heart disease->I50 Heart failure->I63 Cerebral infarction->J18 Pneumonia; unspecified organism
334	AMI Acute Myocard Infr->I25 Chronic ischemic heart disease->I50 Heart failure->J18 Pneumonia; unspecified organism->J69 Pneumonitis due to solids and liquids
335	AMI Acute Myocard Infr->I25 Chronic ischemic heart disease->I50 Heart failure->J18 Pneumonia; unspecified organism->A41 Other sepsis
336	AMI Acute Myocard Infr->I25 Chronic ischemic heart disease->I50 Heart failure->K55 Vascular disorders of intestine
337	AMI Acute Myocard Infr->I25 Chronic ischemic heart disease->I48 Atrial fibrillation and flutter->I50 Heart failure->N17 Acute kidney failure

338	AMI Acute Myocard Infr->I25 Chronic ischemic heart disease->I48 Atrial fibrillation and flutter->I63 Cerebral infarction->J18 Pneumonia; unspecified organism
339	AMI Acute Myocard Infr->I25 Chronic ischemic heart disease->I48 Atrial fibrillation and flutter->I63 Cerebral infarction->I64 Stroke, not specified as haemorrhage or infarction
340	AMI Acute Myocard Infr->I25 Chronic ischemic heart disease->I48 Atrial fibrillation and flutter->J90 Pleural effusion; not elsewhere classified->I50 Heart failure
341	AMI Acute Myocard Infr->I25 Chronic ischemic heart disease->I48 Atrial fibrillation and flutter->I64 Stroke, not specified as haemorrhage or infarction
342	AMI Acute Myocard Infr->I25 Chronic ischemic heart disease->J90 Pleural effusion; not elsewhere classified->I50 Heart failure->J18 Pneumonia; unspecified organism
343	AMI Acute Myocard Infr->I25 Chronic ischemic heart disease->J90 Pleural effusion; not elsewhere classified->I50 Heart failure->N17 Acute kidney failure
344	AMI Acute Myocard Infr->I25 Chronic ischemic heart disease->J90 Pleural effusion; not elsewhere classified->C34 Malignant neoplasm of bronchus and lung
345	AMI Acute Myocard Infr->I25 Chronic ischemic heart disease->I20 Angina pectoris->I50 Heart failure->N17 Acute kidney failure
346	AMI Acute Myocard Infr->I25 Chronic ischemic heart disease->I20 Angina pectoris->I48 Atrial fibrillation and flutter->J90 Pleural effusion; not elsewhere classified
347	AMI Acute Myocard Infr->I25 Chronic ischemic heart disease->I20 Angina pectoris->I48 Atrial fibrillation and flutter->I47 Paroxysmal tachycardia
348	AMI Acute Myocard Infr->I25 Chronic ischemic heart disease->I20 Angina pectoris->I24 Other acute ischemic heart diseases->I50 Heart failure
349	AMI Acute Myocard Infr->I25 Chronic ischemic heart disease->I20 Angina pectoris->I44 Atrioventricular and left bundle-branch block->I50 Heart failure
350	AMI Acute Myocard Infr->I25 Chronic ischemic heart disease->I20 Angina pectoris->K80 Cholelithiasis->K83 Other diseases of biliary tract
351	AMI Acute Myocard Infr->I25 Chronic ischemic heart disease->I20 Angina pectoris->K80 Cholelithiasis->M79 Oth and unsp soft tissue disorders; not elsewhere classified
352	AMI Acute Myocard Infr->I25 Chronic ischemic heart disease->K21 Gastro-esophageal reflux disease->M79 Oth and unsp soft tissue disorders; not elsewhere classified
353	AMI Acute Myocard Infr->I25 Chronic ischemic heart disease->I24 Other acute ischemic heart diseases->I48 Atrial fibrillation and flutter->M25 Other joint disorder; not elsewhere classified
354	AMI Acute Myocard Infr->I25 Chronic ischemic heart disease->I44 Atrioventricular and left bundle-branch block->I47 Paroxysmal tachycardia
355	AMI Acute Myocard Infr->I25 Chronic ischemic heart disease->I44 Atrioventricular and left bundle-branch block->I50 Heart failure->J18 Pneumonia; unspecified organism
356	AMI Acute Myocard Infr->I25 Chronic ischemic heart disease->I49 Other cardiac arrhythmias->I63 Cerebral infarction
357	AMI Acute Myocard Infr->I50 Heart failure->N17 Acute kidney failure->E87 Other disorders of fluid; electrolyte and acid-base balance->J18 Pneumonia; unspecified organism
358	AMI Acute Myocard Infr->I50 Heart failure->N17 Acute kidney failure->E83 Disorders of mineral metabolism
359	AMI Acute Myocard Infr->I50 Heart failure->J18 Pneumonia; unspecified organism->J15 Bacterial pneumonia; not elsewhere classified
360	AMI Acute Myocard Infr->I50 Heart failure->J18 Pneumonia; unspecified organism->A04 Other bacterial intestinal infections
361	AMI Acute Myocard Infr->I50 Heart failure->I63 Cerebral infarction->G40 Epilepsy and recurrent seizures
362	AMI Acute Myocard Infr->I50 Heart failure->I63 Cerebral infarction->J69 Pneumonitis due to solids and liquids

363	AMI Acute Myocard Infr->I50 Heart failure->E11 Type 2 diabetes mellitus->E10 Type 1 diabetes mellitus
364	AMI Acute Myocard Infr->I63 Cerebral infarction->E86 Volume depletion->J69 Pneumonitis due to solids and liquids
365	AMI Acute Myocard Infr->I63 Cerebral infarction->N39 Other disorders of urinary system->N17 Acute kidney failure->J18 Pneumonia; unspecified organism
366	AMI Acute Myocard Infr->I63 Cerebral infarction->N39 Other disorders of urinary system->F05 Delirium due to known physiological condition->J18 Pneumonia; unspecified organism
367	AMI Acute Myocard Infr->I63 Cerebral infarction->I64 Stroke, not specified as haemorrhage or infarction->N39 Other disorders of urinary system
368	AMI Acute Myocard Infr->I63 Cerebral infarction->I62 Other and unspecified nontraumatic intracranial hemorrhage->J18 Pneumonia; unspecified organism
369	AMI Acute Myocard Infr->I47 Paroxysmal tachycardia->I95 Hypotension->N39 Other disorders of urinary system
370	AMI Acute Myocard Infr->K21 Gastro-esophageal reflux disease->K80 Cholelithiasis->K85 Acute pancreatitis
371	AMI Acute Myocard Infr->I71 Aortic aneurysm and dissection->K43 Ventral hernia
372	AMI Acute Myocard Infr->I71 Aortic aneurysm and dissection->J90 Pleural effusion; not elsewhere classified
373	AMI Acute Myocard Infr->I71 Aortic aneurysm and dissection->D64 Other anemias
374	AMI Acute Myocard Infr->I71 Aortic aneurysm and dissection->I63 Cerebral infarction->J18 Pneumonia; unspecified organism
375	AMI Acute Myocard Infr->I71 Aortic aneurysm and dissection->K59 Other functional intestinal disorders
376	AMI Acute Myocard Infr->I71 Aortic aneurysm and dissection->K55 Vascular disorders of intestine
377	AMI Acute Myocard Infr->I35 Nonrheumatic aortic valve disorders->I50 Heart failure->N17 Acute kidney failure
378	AMI Acute Myocard Infr->K25 Gastric ulcer->I50 Heart failure->J18 Pneumonia; unspecified organism
379	AMI Acute Myocard Infr->K25 Gastric ulcer->I63 Cerebral infarction->J18 Pneumonia; unspecified organism
380	AMI Acute Myocard Infr->K25 Gastric ulcer->I35 Nonrheumatic aortic valve disorders
381	AMI Acute Myocard Infr->I95 Hypotension->M25 Other joint disorder; not elsewhere classified->N39 Other disorders of urinary system
382	AMI Acute Myocard Infr->I95 Hypotension->M25 Other joint disorder; not elsewhere classified->J18 Pneumonia; unspecified organism
383	AMI Acute Myocard Infr->I95 Hypotension->J22 Unspecified acute lower respiratory infection->I50 Heart failure
384	AMI Acute Myocard Infr->I95 Hypotension->K59 Other functional intestinal disorders->N39 Other disorders of urinary system

Appendix E
Required documents to access data

Following are the documents required to access research data for this study.

E.1 The HES-APC Database Access

To access the HES-APC database stored in SEED, I was included in the research team led by Dr. Marlous Hall. The research team studied "Hospitalisation and mortality after Acute Myocardial Infarction". The cover letter upon approval is presented below.

NHS Digital

23rd March 2017

NIC reference: NIC-17649-G0X4B

HESID pseudonymisation reference: DLS0684

University of Leeds

Dear Madam/Sir

Accompanying this letter are the files containing the linked data you have requested. All data in this release is subject to the terms and conditions outlined in the data sharing/re-use agreement noted at the top of this document. These are provided as pipe-delimited text files with column headers.

File Name	Year	Row Count
NIC17649_APC_0809	2008 - 09	13,009,872
NIC17649_APC_0910	2009 - 10	15,037,887
NIC17649_APC_1011	2010 - 11	15,835,515
NIC17649_APC_1112	2011 - 12	16,242,593
NIC17649_APC_1213	2012 - 13	16,610,561
NIC17649_APC_1314	2013 - 14	17,175,190
NIC17649_APC_1415	2014 - 15	17,834,991
NIC17649_APC_1516	2015 - 16	18,432,662
NIC17649_APC_1617	2016 - 17	15,733,889

Data provided to you for this dissemination has been classed as out-of-scope for type-2 opt-outs and therefore no records have been withheld from your datasets
NHS Digital has published information regarding Care Information Choices which provides additional information regarding the distribution of type 2 opt-outs.

Please note that we only keep your cohort data for 3 months after we make the linked data available to you. If you require us to keep the data for longer you must make this request in writing with evidence of any necessary approvals for us to do so.

Note that the data will not be identical to that received from the data providers. It has been cleaned by the HES system and in some cases, the "unknown" and "inapplicable" values have been changed to conform to the "typing" requirements of the system. The HES data dictionary explains the meaning of the fields, in particular giving the "unknown and "inapplicable" codes for many fields. The HES data dictionary can be accessed via the HES web pages at http://www.hscic.gov.uk/hesdatadictionary

To un-zip the files provided, use WinZip (version 14.5 or later). The Windows compression tool may cause errors when extracting larger files.

Should you discover that this data does not quite meet your needs you can request an amendment at a slightly reduced rate within 1 month of us making the data available to you. After that time any changes will be charged at the full rate. An amendment is

E.2 SEED Confidentiality Agreement

11 Appendix 1 – Confidentiality Agreement

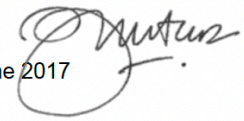
The following is a statement that is to be read and signed, after reading and understanding the document entitled: "Secure Electronic Environment for Data Information Governance Policy". Anyone requiring access to an information asset held in the SEED system must sign the statement before being authorised to access that asset.

AS A COMPUTER USER AT THE UNIVERSITY OF LEEDS I AM AWARE OF THE RULES THAT HAVE BEEN SET OUT CONCERNING THE USE OF THE SEED SYSTEM, AND I AGREE TO COMPLY WITH THESE RULES.

I UNDERSTAND THAT FAILURE TO COMPLY WITH THESE RULES WILL RESULT IN ACTION BEING TAKEN IN LINE WITH UNIVERSITY PROCEDURES

NAME: Guntur P. Kusuma

POSITION: Ph.D research student

SIGNATURE: 

DATE: 14 June 2017

To confirm, below is the Information Governance Lead's signature:

NAME :

POSITION :

SIGNATURE :

DATE :