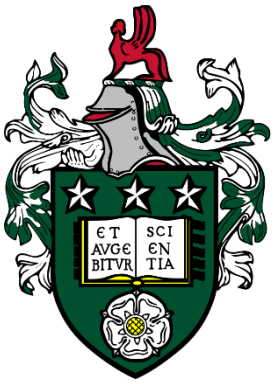
An Analysis of Frailty Progression in Elderly Patients using Process Mining and Machine Learning

Nik Fatinah Binti N. Mohd Farid

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



The University of Leeds

School of Computing

February 2023

Acknowledgement

Alhamdullilah, I praise to Allah my dear God in supporting and allowing me to experience the sweetness and bitterness of this PhD journey. I would like to express my deepest and heartfelt gratitude to my PhD supervisors, Mr Owen Johnson and Dr Marc de Kamps for their priceless and valuable advices, continuous support and encouraging comments. Their critical eyes have been the main contributing factors in pushing forward this research work. I also would like to thank my sponsor, the Ministry of Higher Education Malaysia in giving me a chance to further my study in the UK. Without their financial help, I would never have the experience of studying abroad.

No words can describe how grateful and thankful I am towards the endless prayers and support I received from my beloved mother and father. My much-loved siblings, especially Nik Fatihah who always supporting me whenever the hardest and difficult time came. To Hanis Syazwani my partner through thick and thin time while we were in the UK, I am so thankful for your presence and support. Not to forget, the singLeeds ladies who always together in exploring this marvelous PhD journey.

I am very fortunate to be given an opportunity to work and experience this journey within the process mining healthcare research team lead by Mr Owen Johnson. Thank you to Angelina Kurniati, Guntur Kusuma, Samantha Skyes and Amirah Alharbi for sharing their experiences working in process mining field and giving boundless ideas. I also would like to thank Bradford Institute of Health Research in giving me opportunity and guidance in working with their great team and endless support in working with healthcare data.

Abstract

Frailty is a geriatric medical condition which affect 26% of people with age over 85 in the UK. This distinctive health state happened as a result of cumulative deterioration in bodily systems and diminished clinical state functional reserve over lifetime. It is a transition state from healthy ageing to dependent elderly life. The transitioning happened as the elderly with frailty has low ability to cope with acute illnesses or daily stressors. Understanding frailty progression as part of early identification of frailty may offer great opportunity in lengthening the transition time and maximize healthy ageing period. Majority of user in the healthcare sector identified as the elderly. The abundance of data recorded in the electronic healthcare record (EHR) related to patient health status and conditions has potential to facilitate the understanding of frailty. This thesis used machine learning and process mining techniques which is an emerging data-driven analytic approaches to understand frailty progression, its association with three frailty deficits of concern namely fall, hypertension and polypharmacy and investigate the variation between two frailty scores cut-off points. The main objective in this work is to analyse frail elderly pathway with respect to frailty progression pathway using electronic frailty index (eFI) score and highlight the feasibility of employing process mining and machine learning in analysing frailty. This work used two real-life healthcare datasets to understand the variability of frailty progression pathway. The first dataset is a publicly open healthcare dataset from the tertiary hospital setting in the USA. The first dataset provides as a platform for preliminary work and develop methods in analysing frailty trajectories to support reproducibility of the work. The second dataset is a UK healthcare dataset from the primary care setting. The experiments in the second dataset used as an improvement to comprehensively study frailty progression from the previous dataset. Furthermore, the variation between the literature and proposed cut-off points in this study was investigated using the second dataset. Frailty progression pathway analysis used process mining technique while the identifying the cut-off points in the proposed approach used machine learning technique. The advantages of using process mining in healthcare domain especially in frail elderly has been proven. It is useful in presenting the visualization of frailty progression at different frailty stages and discover the variation of frailty progression with the associated deficits of concern. This thesis also explores the comparison in frailty progression between two cut-off points approaches used in the literature and the proposed approach in this study.

Table of Contents

List of Figures

List of Tables

Chapter 1

Introduction

This chapter will introduce the research background, including the justification on the considerable amount of the data in EHR, which has potential for research and significance of the work within the frail elderly domain area. This chapter is crucial to establish the need for investigating the domain area using a combination of data and process centred approaches, especially process mining.

1.1 Overview of Research

A hospital information system (HIS) is a multidimensional system to facilitate patients-healthcare professionals communication and health-related issues or impairments (Haux, 2006). Its implementation in the healthcare setting contains an immense amount of information. It comes from various sources, including patient medical records, hospital records, results of medical examination, administration records and healthcare devices records. It is essential in every level of healthcare settings; in the first consultation between patient and healthcare professional (primary care), critical care acquiring skilled professionals (secondary care), advanced medical treatment and investigations (tertiary care) and in unusual surgical procedures and diagnostic (quaternary care).

The use of Electronic Health Record (EHR) has grown in popularity and accessible over the last decade. The aim is to bring the healthcare sector towards improved clinicians-patient communication, patients care, safety and efficiency (Menachemi and Collum, 2011). The purpose of EHR is to store and capture the information within HIS. The information varies from clinical or medical data such as laboratory examination or imaging report and patient's medical history such as diagnoses, symptoms, treatments, and prescriptions. One of the benefits offered is that people with some chronic disease who has regular interactions with the healthcare provider will almost certainly have several records, allowing the variability of measurements to be monitored.

Sixty-five per cent of *elderly* people aged 65 years above and living with multiple chronic diseases (Wolff, Starfield and Anderson, 2002). Ten per cent of elderly people are common to have *frailty* which is a geriatric condition identified by eFi

score (Clegg *et al.*, 2013). They will have a large amount of health information recorded that consists of medical conditions related to their diseases and clinical observations. Clinical observations is an act of making clinical judgement such as questioning, measuring and evaluating specimens (Russler, 2009). Healthcare professionals make it to determine the diagnosis, which is part of the medical conditions and construct a treatment plan. The flow of the records in sequence will create a *pathway* to illustrate human disease's mechanism based on the diagnosis and clinical observations.

High variability and an immeasurable amount of data in EHRs offer an opportunity to seek insight. This is especially true given the vast volume of data gathered during healthcare professional and patient interaction. Data analysis helps gain a better view of disease-based diagnosis formation, the varying disease flow and effect of involvement for the proper health outcomes and the potential prospect for change. The trends in data-driven approaches have been getting attention in the last few decades. *Process mining* is an emerging data-driven approach that fills the gap between data and process science. It aims to discovering, monitoring, and enhancing the analysed process from the event log as the input (Van der Aalst, 2016). The input extracted from the structured EHR data to generate a process model. The application of process mining has been made not only in healthcare but in many other areas, including auditing (Jans, Alles and Vasarhelyi, 2014), accounting (Jans *et al.*, 2011) and education (Bannert, Reimann and Sonnenberg, 2014). In comparison, *machine learning* is a learning statistical model method utilising observed data as variables to forecast results or categorise findings (Shalev-Shwartz, Shai, 2014). EHR offer a significant number of variables to enable machine learning technique implementation.

This thesis attempt to apply data-driven approaches such as process mining and machine learning using the EHRs data. The main aim of this study is to explore the feasibility of applying process mining and eFi to the elderly population. The researcher is motivated to help the applied healthcare process mining community and researchers aware of the pertinence of analysing the frail elderly population using process mining. It focuses on the elderly population and will generate a method in understanding the frailty mechanism. Two datasets will be examined retrospectively (1) the MIMIC-III dataset from the tertiary hospital as preliminary work and (2) Connected Bradford dataset from the primary care setting as the main dataset for the study.

1.2 The Elderly and Ageing Population

It is estimated that the population of the elderly is double by the year 2050 from 962 million in the year 2017 to 2.1 billion representing almost 16% of the world's population, according to (Eendebak and Organization, 2015). The majority of elderly people aged 65 or older currently are in the less developed country compared to more advanced age elderly located in a more developed country. The different trend in changing of age composition could be observed in a more developed country; for instance, France took about 100 years to increase the elderly population from 7% to 14% whereas, the rapid pace of change took only a single generation in a less developed country (Suzman, Beard and Organization, 2011).

This extraordinary phenomenon is steered mainly by the decline in fertility rate and advancement in longevity (United Nations, 2017). In the more developed and high-income country, the increasing life expectancy due to longevity, combined with the decreasing in the fertility rate contributing to the increment of the older people aged 60, has made a general trend towards the rapid pace of the ageing population in the world. Moreover, a positive and favourable lifestyle attitude for older people, especially at an advanced age, has proven to be one of the factors of increasing lifespan that promote the reduction of the morbidity level (Rizzuto *et al.*, 2012) (Pandey, 2018).

The longer lifespan provides a tremendous resource and opportunity to explore what older age life might offer (Beard and Bloom, 2015). For example, venture into a new career, continuing in new hobbies or even pursuing education. However, the critical element that will determine the success of achieving those opportunities is health. Nevertheless, there is limited evidence showing that extended period of life come together with good health (Crimmins and Beltran-Sanchez, 2011). However, findings from many studies done in this area are very conflicting and vary in many aspects such as environment modification, assisting devices or increase in physical activities (Manton, Gu and Lamb, 2006; Jagger *et al.*, 2008; Stewart, Cutler and Rosen, 2013).

The extended years joined with lousy health in older people will introduce very negative implications to social and economic. The majority of the elderly are linked with the variable of health conditions and normal functional life developments (Stöber *et al.*, 2015). As for the social implication, these population will affect the supports ratio and the resources of a country by changing the dynamic of policies making such as healthcare services, institutional care,

employment and retirement (Tinker, 2002). Furthermore, many evidence shows that the cost of providing healthcare services is escalating for the ageing population (Palangkaraya and Yong, 2009; Shrivastava, Shrivastava and Ramasamy, 2013; Ilinca and Calciolari, 2015).

1.3 Frailty and its Impact

The elderly population has high exposure to developing frailty, especially those who suffer from comorbidities or multiple chronic diseases (Hoogendijk *et al.*, 2019). *Frailty* describes an individual who is in poor health and has a high vulnerability to the impact caused by the internal and external stressors events (Rockwood, 2005; Hogan, 2018). It is a common clinical syndrome among the elderly, which is caused by the cumulative decline in several organs function over time (Clegg *et al.*, 2013). It also recognised to have multidimensional syndromes such as physical, psychological and social (Sieber, 2017).

Many studies found that the frail elderly cannot adapt to a stressor such as mild infection and trauma. They are strongly linked to various unfavourable health effects, including institutionalisation, disability, lower quality of life and mortality (Fried *et al.*, 2001; Walston *et al.*, 2006). It aligned with considerable evidence shows that the majority user of healthcare and social services in the UK and global is the elderly making the cost of healthcare intensified (Oliver, 2009). In the United States alone, for example, the top 10% of their population utilised 42.5% of healthcare expenditure by the elderly aged 65 or older (Zayas *et al.*, 2016). In addition to that, according to several studies, the cost of healthcare for frail elderly is many time higher than their non-frail elderly counterparts (Bock *et al.*, 2016; Simpson *et al.*, 2018; Salinas-Rodríguez *et al.*, 2019).

1.3.1 The Course of Elderly towards Frailty

The journey of care for frail elderly involves several services. Various healthcare professional involvement such as GPs, pharmacists, geriatricians, nurses, social care workers and other specialists are required within the services. The six fundamental elements of healthcare provision for the elderly are 1) primary care, the first point of contact with healthcare provider 2) urgent care, a patient who needs urgent medical attention but not severe enough to go to the emergency department 3) intermediate care, exhaustive coverage of services to prevent unnecessary admission to hospital 4) acute care or secondary care 5) chronic

disease management and 6) end of life care (Ayyar *et al.*, 2010). The elderly may require one element at one time in a continuous process or multiple at one time.

The course of the elderly towards frailty can be defined using the clinical frailty scale components widely used to evaluate frailty based on clinician judgement (Mendiratta and Latif, 2021). It consists of nine scales: 1) very fit, 2) well, 3) managing well, 4) vulnerable, 5) mildly frail, 6) moderately frail, 7) severely frail, 8) very severely frail, and 9) terminally ill. The clinical frailty scale is one of the assessments to identify frailty and to require clinical judgement. Alternatively, the electronic frailty index score (eFi) is a frailty assessment tool developed to assist clinicians in stratifying the severity of frailty directly from the EHR without requiring any clinical judgement (Clegg *et al.*, 2016).

It is a dynamic state and long-term condition which can deteriorate or improve (Puts *et al.*, 2017; Travers *et al.*, 2019). However, when frailty starts to develop and progress naturally, other geriatric syndromes such as accelerated physical deterioration, pressure ulcers, incontinence, fall and cognitive problems are following later (Xue, 2011).

1.4 Exploiting EHR data using Data-Driven and Process-based Approach

The study to recognise frailty condition has been a significant focus in the clinical domain aim to distinguish between frail and non-frail elderly (Dent, Kowal and Hoogendijk, 2016). It allows healthcare professionals to target the best possible treatment to avoid any harmful medication and invasive procedure, putting them at more risk. However, chronological age is not a generalisable indicator of health status and identifying vulnerability in the elderly is challenging (Finkel, Whitfield and McGue, 1995; Romero-Ortuno and Kenny, 2012). The reason being that in reality, as people get older, their health status becomes more diverse.

While EHRs are primarily used for clinical purposes, they have also been used for research as they represent longitudinal data routinely collected during healthcare delivery. It opens the door to research as it can answer questions that are difficult to answer using randomised clinical trials or traditional cohort studies with data collection. The work they perform ranging from biomedical studies, observational studies and epidemiologic studies of cross-sectional studies within a single hospital and longitudinal studies on geographically dispersed patients (Cowie *et al.*, 2017). Apart from that, the data analytics trends have been made in the last decades and offer approaches to find interesting insights from the

routinely collected data. Process mining and machine learning are some of the many approaches that have excellent capabilities in revealing fascinating information within data (van der Aalst, 2011; Kim *et al.*, 2019).

Process mining was initially proposed to improve the workflow or pathway of the management system by (Van Der Aalst, Weijters and Maruster, 2004). It is rather challenging as the traditional approach to creating a workflow requires a deep understanding of the system from the involvement of various management personnel and workflow designers. Following that, the process mining technique aims at discovering the workflow or pathway model simply from the event log produced from any transactional system. Not only process mining capable of creating a model to reveal the control-flow perspective but also capable of analysing from other perspectives such as data, organisational and performance of event log (Van der Aalst and Weijters, 2004).

Apart from that, machine learning is an intersection between statistical and informatics which considerably associated with knowledge discovery and data science (Jordan and Mitchell, 2015). It applies statistical models to find patterns from the complex healthcare data and formulate hypotheses that are not entirely based on the assumptions of data distribution (Beam and Kohane, 2018). It is formed as an essential part of incorporating machine learning into EHRs as the primary healthcare provider involves hypothesis generation, testing and actions. Although both process mining and machine learning could be used for hypothesis testing analysis, different technique served different purposes during analysis.

1.5 Problem Statement

Frailty is undoubtedly a serious issue that should be a significant concern given most healthcare burdens coming from the minor world population. It is proven following vast works on frailty been done and summarise using systematic review; care interventions within the primary care (Travers *et al.*, 2019), improving patient handovers (Hesselink *et al.*, 2012), randomised control on the most effective interventions (Kidd *et al.*, 2019) and even screening tools for frailty in primary care (Pialoux, Goyard and Lesourd, 2012). In addition to that, works on frailty in elderly people within the data-driven approach or informatics field also has caught many attentions, such as modelling frailty condition using machine learning (Tarekegn *et al.*, 2020), data mining approach for post-stroke mortality assessment (Easton, Stephens and Angelova, 2014) and text mining approach using EHRs for building physiotherapy corpus within elderly (Delespierre *et al.*,

2017). However, the generalisability of published research on frailty with the implementation of process mining is still limited based on our earlier work summarised in (Farid, De Kamps and Johnson, 2019) and so far only one implementation of process mining looking into behavioural of nutritional assessment within elderly (Valero-Ramon *et al.*, 2019).

Although past works applying process mining within the healthcare domain have shown promising result, the techniques used in diverse within different focused healthcare domain, one of the possible reasons is the complexity of the healthcare record. If it is ignored, a complex process model will be produced, hard to understand and analyse. In contrast, the general approach suggesting to generate several process model from splitting the main event log into sub-parts it still unable to discover the entire behaviour of the event log (van der Aalst, 2011).

1.6 Research Aim, Objectives, and Questions

The main objective of this study is to analyse the frail elderly pathways with respect to frailty progression using eFi. A hypothesis is that process mining can be implemented to analyse frail elderly pathways using EHR. The research questions were established based on the big question on “how can we analyse frail elderly pathways using eFi score?”.

(RQ1) Is it possible to analyse frail elderly pathway using data and process mining analysis?
(RQ2) What is the best illustration to portray frailty pathway?
(RQ3) Is it possible to analyse frailty progression utilizing eFi scores?
(RQ4) How can the dataset in relation to eFi scores be extracted from the EHR system?
(RQ5) Are the cut off points used in eFi literature confirmed by real life data?
(RQ6) Is it possible to determine new cut off points following data-based approach?
(RQ7) What features can be used to characterize the new cut off points in eFi scores?

Seven (7) research questions (RQ) formed to navigate the work in frailty using data and process-based analysis. The first (RQ1) “Is it possible to analyse frail elderly pathway using data and process mining analysis?” is focusing on the feasibility of analysing frail elderly pathway in identifying the best approach for analysis. It includes aspects of pathway in frail elderly that beneficial to

investigate frailty as in (RQ2). Those aspects can be defined based on the process related to frailty pathway. The aspect can be derived from variety of process perspectives representing the process such as the outcome based-perspective, event sequence-based perspective or interval-based perspective. The (RQ3) is explicitly derived from (RQ1) whether “it is possible to analyse frailty progression through eFi scores?”. Nevertheless, the pivotal question rooted on how the extraction of dataset can be done from the EHR system entailing an event log of frailty pathway (RQ4). This is undoubtedly an initial challenge in this work as the database which stored patient records are not logged based on process perceptions, hence obtaining the right dataset for analysis is an essential work.

In addition to that, (RQ5) is the next question “Are the cut off points used in eFi scores literature confirmed by real life data?” acts as the second pivotal part in this work. The work extends by exploring the plausibility of implementing other approach in establishing the cut off points for frailty categories and whether it conform with the reported findings. The confirmatory analysis consists of two separate questions in this research. Firstly, it is investigated through (RQ6) “Is it possible to determine the new cut off points following data-based approach?”. This RQ requires to determine the potential of applying the best data-based approach in identifying the new cut off points. To support this question, the features that fit to characterise the cut off points need to be identified as in (RQ7) “What features can be used to characterise the new cut off points in eFi scores?”.

1.7 Study Approach

The general approach of the study involves several steps, (i) exploration of dataset from the patient records is extracted from the EHR system to discover frailty pathway, (ii) separate analyses will be conducted as experiments comprises of frailty progression and confirmatory analysis for cut off points in determining frailty categories and (iii) evaluation of the findings following domain experts verification and conformation with the past studies. The primary input of this work will be the dataset extracted from EHR database consists of patient diagnosis, clinical observations, medication prescriptions and administration information. The extracted dataset were transformed using process mining and data analytical approaches to represent the frailty pathway to understand frailty progression in illustrating model that easy to understand by the domain experts. The output of the analysis will be presented in various visualisations to help in understanding the frail pathways and its progression.

The methodology of the study adopted from the widely-used methodology in the process mining research area, which is process mining project methodology (PM2) (Van Eck *et al.*, 2015). The methodology will be discussed in detail in Chapter 3. The first phase of the methodology starts with (i) planning to establish the aim, research questions and scope of the analysis. Next is (ii) extraction phase is the identification of suitable dataset extracted from the patient records database. It followed by (iii) data processing and transformation where the raw data extracted from the database is cleaned and converted to characterise frailty pathway and its progression. The essential phase is (iv) mining and analysis, where the analysis of frailty pathway and its progression is conducted. Finally, the last phase (v) evaluation was done following the verification from the domain experts on the findings discovered in analysis as well as through past studies verification.

Two datasets were used in the study to investigate the frailty pathway using the routinely collected records from EHR system. The source of the dataset came from (i) hospital in tertiary care setting in the United States of America (USA) and hospital in primary care setting in the United Kingdom (UK). The main differences of these two sources of dataset is that USA healthcare system is dominant by the private sector while UK is by government sector (Adeniran, 2004). This high variability within two datasets create an opportunity for an extra analysis of comparison between two countries internationally. The datasets will be explain in detail in Chapter 4 for USA dataset and Chapter 5 for UK dataset.

1.8 Publications and Presentation of This Work

- Paper 1: Farid, N.F., de Kamps, M. and Johnson, O.A., 2019. Process mining in frail elderly care: a literature review. In *Proceeding of the 12th International Joint Conference on Biomedical Engineering Systems and Technologies-Volume 5: HEALTHINF* (Vol. 5, pp. 332-339), SciTePress, Science and Technology Publications.
    - The work in this paper will be discussed in detail in Chapter 2 Section 2.2.1.1.
- Paper 2: Farid, N.F., de Kamps, M. and Johnson, O.A., 2022. A Process Cube Based Approach of Process Mining in Analysing Frailty Progression Exploiting Electronic Frailty Index. In *Proceeding of the 15th International Joint Conference on Biomedical Engineering Systems and Technologies-*

Volume 5: HEALTHINF (Vol. 5, pp. 605-613), SciTePress, Science and Technology Publications.

  - The work in this paper is part of Experiment 5 conducted from Chapter 6 in Section 6.4.4.1.

- Presentation 1: Farid, N.F., David Mehdizadeh, de Kamps, M. & Johnson, O.A., 2021. Modelling the Progression of Frailty in Elderly People using Process Mining and the Electronic Frailty Index (eFI), presented in Process-Oriented Data Science for Healthcare (PODS4H).
  - The work in this paper is part of Experiment 5 from Chapter 6 in Section 6.4.4.

1.9 The Structure of the Thesis

The thesis is structured following the experiments conducted in analysing the frail pathway and progression. The work first started with preliminary analysis using the MIMIC-III dataset, and next dataset in this work is from the Connected Bradford dataset. The structure of the thesis is illustrated in Figure 1.1. The box linked by blue arrows indicate the flow of the thesis chapters and its connection to each research questions. It comprises eight (8) chapters with three analysis chapters. The aim of the involvement of research questions in Chapter 2 is for reviewing past studies. It is for establishing the current relevancy of the existing works related to research questions. The remaining analysis chapters 4, 6 and 7 proceeds with hypothesis testing to answer respective research questions. The description of each chapters are as follows:

Chapter 2: **Literature Review**. In this chapter, each of the domain areas for research is discussed in detail. It summarises the background of healthcare in the frail elderly and the technical area of process mining and data-based approach. The healthcare domain includes frailty mechanisms in the elderly, several tools used to recognise by healthcare professionals. The technical part of process mining will include the background in process mining, various representational process models known as modelling notation and process mining algorithm. Apart from that, the literature that encompasses the combination of healthcare in process mining will be discussed.

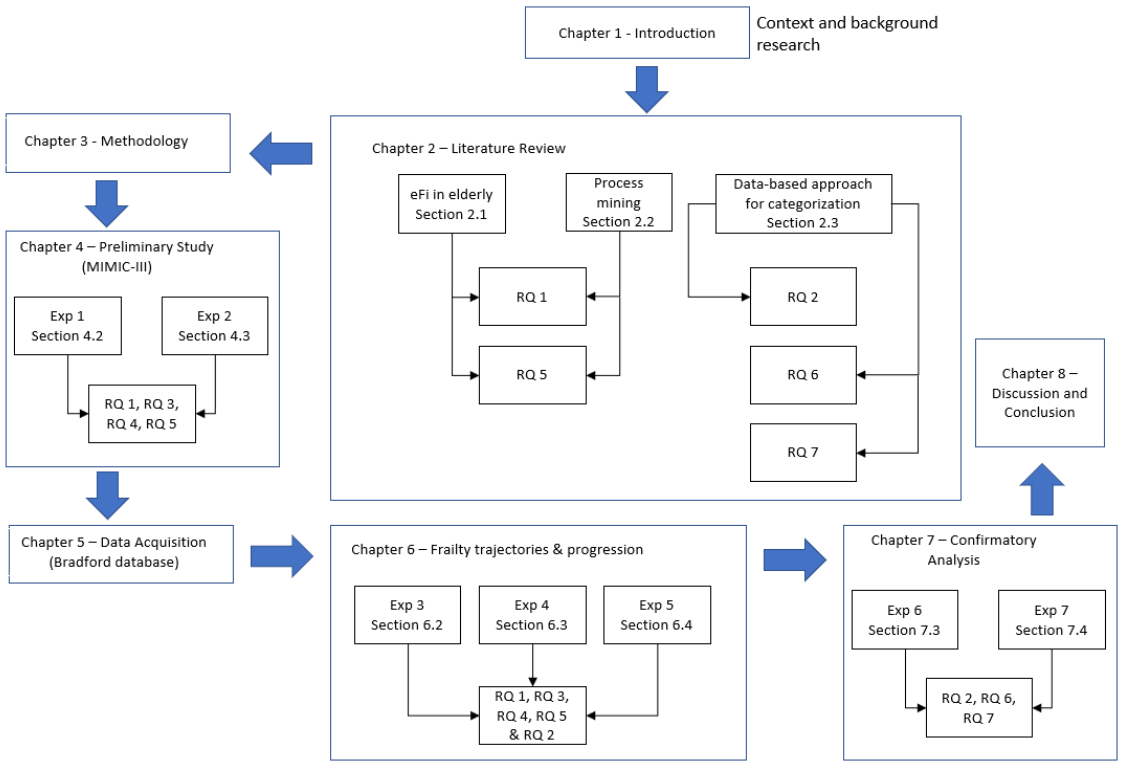


Figure 1. 1 The outline of thesis structure. Each chapters are connected with arrow to illustrates the flow between chapters.

Chapter 3: **Research Methodology**. This chapter explains the methodology implement into the work to conduct the analysis. It follows the PM2 methodology, which covers the 1) planning, 2) extraction of event logs from the raw dataset, 3) data transformation and processing, 4) mining and analysis and 5) evaluation.

Chapter 4: **The Preliminary Study: MIMIC-III Dataset**. This chapter will explore the initial work done using the openly available dataset from the US tertiary healthcare system. It comprises two parts, where the first will provide the overview of the dataset, data characterisation and selection, data cleaning and preparation. The next part will discuss the experiments conducted as the preliminary work in this thesis.

Chapter 5: **Acquisition of Event Log: Connected Bradford Dataset**. This chapter will explore the primary dataset used for the work, directly extracted from the EHR of the primary care setting. It first describes the study setting and context of the extracted dataset, the provenance of the dataset, data characterisation, data cleaning and preparation and lastly, cohort selection and characterisation.

Chapter 6: Frailty Trajectories and Progression. This chapter will describe the experiments conducted (i) to model frailty trajctories and (ii) to examine the progression rate of frailty between different categories and the association between focused deficits and. The focused deficits are fall, hypertension and

polypharmacy and examine the association happened within different frailty stages. This chapter will aim to answer research question three and four in the thesis.

Chapter 7: **Confirmatory Analysis**. This chapter discusses the experiment in investigating the cut-off points of eFi scores. The cut-off points are determined to classify the frailty categories following data-based approach. The following chapter will explore the first experiment conducted after establishing the cut-off points of frailty categories. The experiments will be working on the connected Bradford dataset to answer the work's second research question. The comparison of models between different frailty categories also will be discussed in this chapter.

Chapter 8: **Discussion and Conclusion**. This part of the chapter will explore challenges posed during applying data-driven approaches, process mining using EHR within the frail elderly domain, and the thesis's contribution. This chapter's central part will discuss and conclude findings gathered from all experiments done in the thesis regarding the improvement method applied in work. In this chapter also, the limitation and the potential of future work will be discussed as well.

Chapter 2

Related Work

This section begins with the introduction which includes the definition of frail elderly, the prevalence of frailty among older people; the differences can be observed between fit elderly and frail elderly people and the stages of frailty process will be explained further along with its reaction with stressor event. The following sections will discuss, process mining, application of process mining in healthcare domain along with its tools and methodology. Finally, machine learning and its application in process mining will be briefly explained.

2.1 Frailty in Elderly

According to United Nations population report in 2015 “population ageing is accelerated rapidly worldwide, from 461 million people older than 65 years in 2004 to an estimated 2 billion people in 2050”. In addition to that, it also reported that the number of older people aged 80 and above grows triple in number from 125 million to approximately 5 million people in 2050. Improvement in a medical study declined in child death and rising living standards increase the survival and life expectancy contribute to the increasing number of older people. However, the increase in population raises another challenge to health and social system which is high consumption of healthcare spending and utilisation of resources (Shugarman, Decker & Bercovitz, 2009). Hence, it is crucial to have proper planning and strategy of care and treatment in elderly people.

Ageing population itself is not the real issue to the healthcare system but the relation between ageing and frailty in the elderly. Frailty is common among elderly people; however not all old people are frail and not all frail people are aged. The prevalence of frailty has been reported between 5-58% from the literature review done from 1997 to 2009 by Sternberg, Schwartz & et al., (2011). Moreover, frail elderly people are found to be around 20-30% among older people aged 75 years above (Topinková, 2008). Interest has been growing on frailty among scientist and clinicians since frailty has appeared to be one of the real geriatric condition besides functional declines and disability.

Frailty is described as a situation of high susceptibility to external and internal stressor event which is caused by a cumulative decline in several organs function over time (Clegg, Young & et al., 2013). It also recognised to have multidimensional syndromes such as in physical, psychological and social (Sieber, 2016). The inability to maintain normal cellular body function may result in difficulty in managing with everyday activity independently (Li Xue, 2011) and increase chances of adverse health outcomes such as falls, hospitalisation, institutionalisation, delirium and even mortality (Fried & et al, 2001; Crandall, Duncan & et al., 2016; Mitnitski, Mogilner & et al., 2001; Eeles, White & et al., 2012). Figure 2.1 shows the different effect of stressor event and its relationship with functional abilities across frail and non-frail people. The upper line shows any minor illness attacking non-frail people which result in small changes of decline in functional abilities, recover to the same level in a short amount of time and still living independently. Whereas, there is a significant drop in functional abilities in the bottom line which result in changing from independent to dependent and require a quite a long time to be independent again.

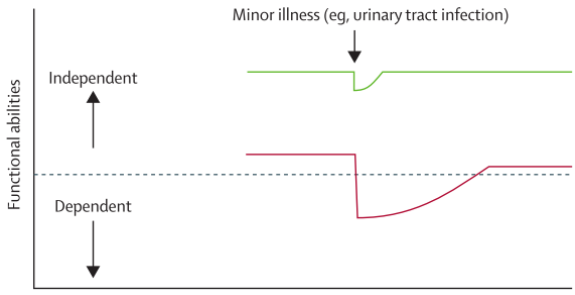


Figure 2. 1 The relationship of stressor events with the functional abilities in (Clegg, Yung & et al., 2013)

There is no firm agreement in the complete and operational definition of frailty on how to best quantity it even though it has been growing in the literature. According to Dent, Kowal & Hoogendijik (2016), there are several reasons as to why it is hard to define frailty accurately: the complex and interrelated cause of it, difficult to differentiate between disabilities caused by frailty or ageing and the individualistic work done by frailty researchers. However, the rule-based definition and frailty index are two popular definitions which would best describe frailty (Conroy & Elliot, 2017) that will be discussed more in the next section.

2.1.1 Frailty Process

The frailty progression seems to be in a transitional state from a good to reduce functional abilities in a dynamic progression as investigated by Gill & et al., (2006). According to Sieber (2016), the progression caused by a variety of factors consists of social, physical and psychological dimensions and not only focus onto pathophysiology condition which is an abnormal biological process resulted from the ageing process. Most studies found that common abnormal biological process associated with frailty are skeletal muscle, endocrine system, inflammatory and nervous system (Waltson, Hadley & et al., 2006). Although, there is evidence that malnutrition and progressive loss of muscle (sarcopenia) also contribute to the cause of frailty development (Calvani, Marini & et al., 2015).

Frailty researchers believe that frailty is a state that falls in between normal ageing, pre-frail state on one end and severe frailty, end of life at the other end of the continuous sequence of frailty process (Lekan, Wallace & et al., 2017). Figure 2.2 below shows the dynamic progression of frailty and its association with performance or functional abilities and homeostatic mechanism or inability of organs to function normally. The first top dotted line represents the cutting point between fit and frail person. The pre-frail state is where a person body could still respond to any stressor event but not in a rigorous way as in the normal ageing state with the possibility of complete recovery. Whereas, when a person in the frail state the full recovery could not be achieved due to the inability of organs function to respond to any stressor event. Next, the complication state is where adverse outcomes of frailty happened such as hospitalisation and institutionalisation frequently, morbidity, disability and even mortality (Sternberg, Schwartz & et al., 2011). However, according to Lekan & et al. (2017), the early phase of frailty is always changing give the opportunity to reverse or prevent the progression of frailty but not at the end phase because of advanced multisystem dysregulation which is hard to control.

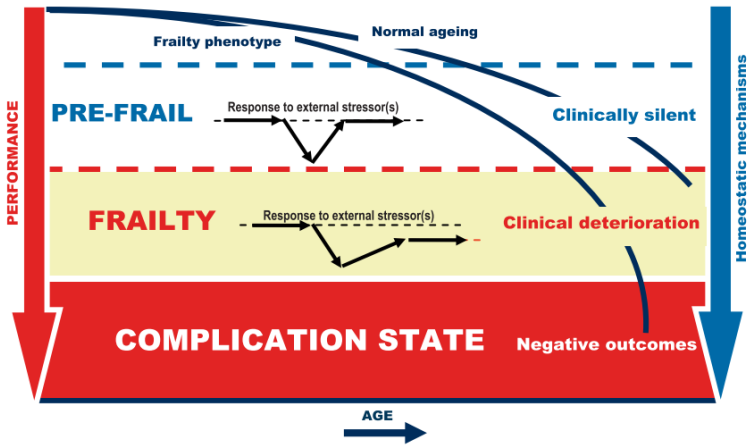


Figure 2. 2 Frailty progression in (Lang, Michel & Zekry, 2009)

The transition of frailty is characterised by the frequent change of state over time (Gill & et al., 2006) it is mostly taken place in one directional from the fit, to pre-frail and to its climax state which is a frail state. There are few studies found on examining the frailty transition in population (Li Xue, 2011; Bentur, Sternberg & Shuldiner, 2016; Gill & et al., 2006) they concluded that the transition is usually happening from less frailty to more frailty than the other way around. However, many studies are investigating the possibility of reversing the transition. It is believed that the improvement of the condition is possible in the situation of better control of criteria that contribute to the development of frailty (Rothman, Summers & et al., 2008).

2.1.2 Frailty Assessment Tools

The multidimensional symptoms of frailty (Sieber, 2016) make it hard to decide which part of the domain should be included as the assessment criteria to identify frailty. Several assessment tools are applied to different domains with different weight and currently, there is still no general agreement achieve. Based on previously reviewed literature the most domain used as the criteria are physical function, gait speed or mobility, and cognition (Fried, Tangen & et al., 2001; Mitnitski, Mogilner & et al., 2001; Turner & Clegg, 2014) with the addition of assessment criteria included nutritional status and mental health. The additional criteria are obtained via a Delphi-based consensus of experts (Conroy, 2017) in order to clear the issue.

There are several frailty assessments available such as PRISMA-7 (Turner & Clegg, 2014), Sherbrooke Postal Questionnaire (SPQ) (Hábert, Bravo & et al., 1996), Gérontopôle frailty screening tools (GFST) (Vellas, Balardy & et al., 2013),

Tilburg Frailty Indicator (TFI) (Gobbens, Assen & et al., 2010), Groningen Frailty Indicator (GFI) (Steverink, Slaets & et al., 2001) and other frailty measurement. Each of the assessment tools will be discussed as follow:

a) PRISMA-7, SPQ and GFST are an example of assessment based on a set of questions to identify frailty. However, the GFST will have clinical judgement regarding frailty status in the second step of assessment. PRISMA-7 with gait speed test and timed-up-and-go test are recommended by the British Geriatric Society (Turner & Clegg, 2014). Presence of frailty could be identified when a person obtains a score of less than 0.8m/s in gait speed, more than 10s in timed-up-and-go test and more than or equal to three in the questionnaire. The questionnaire consists of seven parts which are aged more than 85 years, male, limited in activities due to a health problem, dependent on others, stay at home due to health problems, social support and use of walking aids such as a walker, cane or wheelchair.
b) TFI was developed based on Netherland population during 2010. It consists of three main domains of frailty which are; psychological frailty including cognition, depressive symptoms, anxiety and coping; physical frailty include weight loss, difficulty of walking, health, vision problems, hearing, balance, strength in hands and physical tiredness; and the last domain is social frailty include living alone, social relations and social support. The highest score for TFI is fifteen and frailty is a presence when a score of five or above is obtained.
c) The third frailty assessment is GFI that consists of fifteen questionnaires that cover a variety of domains including mobility, vision, hearing, nutrition, comorbidity, cognition, psychosocial and physical. A frail person would need to score four or greater from this validated frailty assessment (Drubbel, Bleijeberg & et al., 2013).

Although some frailty assessment tools discussed above shows good accuracy in determining frailty and have the predictive ability of adverse outcome, they still lack in terms of practicality. This is because they are used in the population base setting only (Dent, Kowal & Hoogendijik, 2016) but not in clinical. Besides the previous frailty assessment, the emerging frailty assessment models are phenotype model (Fried, Tangen & et al., 2001) and cumulative deficit model (Mitnitski, Mogilner & et al., 2001). These two models have been used both in a clinical and population-based setting and have a better predictive outcome (Dent, Kowal & Hoogendijik, 2016).

2.1.2.1 Phenotypes Model

Phenotype model assesses frailty based on the following criteria: weight loss, exhaustion, weak grip strength, slow walking speeds and low physical activity (Fried, Tangen & et al., 2001) and measure level of degree in each component of criteria. This model categorizes patients into either group of not frail, pre-frail or frail. A patient is considered as frail when he or she possesses three or more positive criteria and not frail where no requirements are present. While pre-frail is defined when there are less than three positive criteria detected. This category is believed to have a high progression of frailty.

2.1.2.2 Cumulative Deficit Model

The second assessment tool is the cumulative deficit model; it does not directly consider any of the criteria found in frailty but concentrating on symptoms, diseases, signs and disabilities as deficits identified in a comprehensive geriatric assessment (Rockwood & Mitnitski, 2007). This mathematical model of frailty counts by summing up the deficit present and divided by all possible deficit identified for frailty. The more deficit present in an individual, the frailer he or she is. However, according to (Rockwood & Mitnitski, 2006) a limit to which the highest deficit could get is 0.67 which is more than the value the likelihood of survival is impossible. The electronic frailty index which is developed based on this model is now embedded in SystemOne since July 2014 and EMIS web (Conroy, 2017) in April 2016.

Although both assessments have been vigorously developed and validated in many literatures (Song & et al., 2010; Clegg & et al., 2016; Bandeen-Roche et al., 2006; Makary et al., 2010), the clinical application of these two models should be deemed differently (Martin & Brighton, 2008). These could be noticed that phenotype model is easier to work and focus on the physical characteristic of frailty, whereas, frailty index model need more intricate work since it involves a more various area of frailty in details. However, both used comprehensive geriatric assessment criteria that contribute to significant functionality in the model.

A) Electronic Frailty Index (eFi) scores

A standard procedure for creating frailty index was introduced in (Searle *et al.*, 2008). Frailty index is a tool developed to identify frailty based on the cumulative deficits model. Its severity was measured by the level accumulation of diseases associated with frailty. In this work, the measurement of electronic frailty index

(eFI) scores was determined using the EHR following the work in Clegg et al. (2016) and discussed in Section 5.7.2 in this thesis.

2.1.3 Frailty Elderly Care Pathway

The possibility of adverse outcomes to happen in frail elderly such as increase medication, repeated falls, hospitalisations, institutionalisation and death is increasing as the development of frailty is growing. Providing a better strategy of treatment care and validated the rehabilitative program able to delay or reduce adverse outcomes (Lang, Michel & Zekry, 2009). Moreover, the author suggests for a quality care processes for a frail elderly patient in several domains of care as aims in slowing down or halting frailty progression (Arora, Johnson & et al., 2007) as they found there was the poor quality of care in the elderly patient as compared to general medical conditions.

2.2 Process Mining

An event log is a transactional records logged into any information system electronically. They composed of information made during any process within business or institution such as the activity made, person who executed the activity, time which the activity has taken place and other additional information related to the activity performed. The abundance of transactional logged provides an opportunity for a posterior analysis.

Process mining is an emerging analysis approach of event-data type focus on extracting information and knowledge hidden about a process and provides a great management tool. It works by analysing a set of events log-built from data generated that stored in the database system. An event log can be regarded as a series of activities taken place that shows the actual behavioural operational process related to a case or process instance.

Basically, there are three main concepts of process mining; discovery analysis where it reveals the actual process happened based on given event logs, conformance analysis and enhancement (van der Aalst, 2011). The overview of three types of process mining is shown in Figure 2.3 below. The purpose of discovery analysis is to discover hidden structure of process and put it in the form of visualisation model of different type (e.g. BPMN model, Petri nets, process trees and statecharts) by capturing the divergence and dynamic behaviour of the actual process, later conformance checking analysis will be applied to conform the model recorded in event log to the recommended model or a guideline of the

process. Whereas, for enhancement analysis used meaningful information in the event log to improve and even make an extension to the process model.

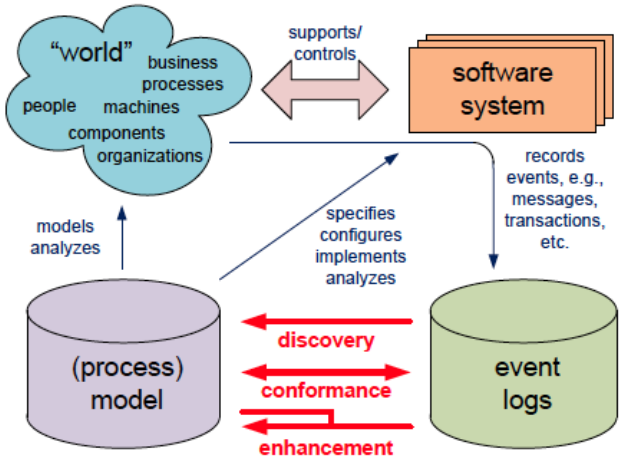


Figure 2. 3 Overview of process mining in (van der Aalst, 2011)

The central analysis perspective of process mining is a control-flow perspective that focuses on the order of activities in the event log. However, process mining is not limited to that perspective since the richness of information contained in event log offer another aspect of perspective to be analysed such as organizational, time, performance and case perspective (van der Aalst, 2011).

2.2.1 Process Mining in Healthcare

Process mining is applicable to a variety of domain and a growing body of literature given the great ability of process mining in extracting valuable information from time-oriented activities and its relationship between resources of the organization. Furthermore, a systematic literature review was done looking at healthcare to investigate further the promising approach of process mining by Rojas, Gama & et al. (2016) identified 22 different case studies in various medical fields applied a variety of process mining techniques.

There are several application of process mining that has been done recently in healthcare such as Garg & Agarwal (2016) analysed the patient flow data obtained from private community hospital, case study of medical service pathway of patient with Type 2 diabetes mellitus (Lismont, Janssens & et al., 2016), improving the layouts of emergency department of non-critical and critical patients route (Rismanchian & Lee, 2016) identifying differences in process care of chest pain management in four hospitals in Australia using process mining techniques; discovery and conformance analysis, and looked at the control flow

and performance perspectives (Partington, Wynn & et al., 2015) and improving waiting time of queuing service in healthcare (Yampaka & Chongstitvatana, 2016).

Process discovery technique was reported to be used in all study where only one applied additional conformance analysis (Partington, Wynn & et al., 2015) from the studies listed above. Whereas in process mining perspective, all studies focused on control flow perspective with only Lismont, Janssens & et al., (2016) had model represents an organizational perspective. The list of examples of using process mining in healthcare showed that process mining is applicable in healthcare domain and could be extended whether in specific healthcare domain or in general. The process mining techniques could be explored in all perspective based on the problem that wants to be solved.

2.2.1.1 Process Mining in Frail Elderly

As part of this study, a systematic literature review of previous studies implementing process mining has been published [My SLR] and section below will summarised the paper.

The systematic literature review has been conducted in June 2018 using the keyword search and following the review process adopted by Kurniati *et al*, (2016):

(“Process Mining” OR “Workflow mining” OR “Pathway mining”) AND (“Frailty” OR “Elderly” OR “Older adults” OR “Ageing” OR “Geriatric” OR “Palliative” OR “Debility” OR “Decrepit” OR “Deteriorate” OR “Vulnerable” OR “Senile” OR “Impairment” OR “Fallibility” OR “Senescence”)

Three screening review processes to carefully selected papers were done from a total 1,091. The papers were retrieved from variety of sources in Google Scholar, PubMed, BMJ Open, ScienceDirect, Elsevier/Springer, ACM, Web of Science, Medline, DBLP and processmining.org. The screening steps are based on title-based checking, abstract-based checking and content-based checking with inclusion criteria of (i) English articles published from 1998, (ii) peer-reviewed or conference proceeding articles, (iii) article with process mining case studies in frail elderly domain, (iv) no duplication and (v) only articles and not book. As a result, eight papers were selected with five themes identified, (i) data and process type, (ii) geographic analysis, (iii) methodology, (iv) medical settings and (v)

challenges faced when conducting the studies. A summary of the result of each theme are discussed in the following paragraph.

The most common papers used the type of data source extracted from the administrative healthcare system within the organizational process. Four papers gathered data from sensors in smart living environment to study elderly behaviours (Vitali & Pernici, 2015; Tax *et al*, 2018), one from smart home (Tapia *et al*, 2004) and from nursing homes (Llatas *et al*, 2011). The other two papers used data of elderly who in needs of ambulant services are Wolf *et al.,* (2013) and Munstermann *et al*., (2012). Two papers used healthcare data collected from the EHR of acute care and surgery Najjar *et al*, (2018) and EHR administrative data Conca *et al,* (2018). The six papers investigated the organizational process of elderly daily activities in a controlled environment as the process type for the study (Vitali & Pernici, 2015; Tax *et al*, 2018; Tapia *et al*, 2004; Llatas *et al*, 2011; Wolf *et al.,* 2013; Munstermann *et al*., 2012). Meanwhile, Conca *et al*., (2018) used the elderly data with Type 2 Diabetes to examine the different interaction of healthcare professional roles and one paper studied the clinical treatment process (Najjar *et al*., 2018).

Europe is the most common country where the data was retrieved for analysis. Two papers obtained the data from Germany (Wolf *et al*., 2013; Munstermann *et al*., 2012). While Triki et al., (2015) from France, Vitali & Pernici, (2015) from Italy, Tax *et al*., (2018) from Netherlands, Lltas *et al*., (2018) from Spain, Conca *et al*., (2018) from Chile and Najjar *et al*., (2018) from Canada.

None of the papers appeared to implement the process mining methodology of L* life cycle nor the process mining project methodology. It was apparent that the papers had established their own methodology in applying process mining analysis.

Different medical domains were explored where two papers examined the progression of dementia in elderly (Lltas *et al*., 2018; Wolf *et al*., 2013), elderly suffered from the heart disease was studied in Najjar *et al*., (2018) and Conca *et al*., (2018) focus on the elderly with Type 2 Diabetes mellitus. Meanwhile, the other four papers did not mention the specific medical domain associated with their study (Triki et al., 2015; Tax *et al*., 2018; Vitali & Pernici, 2015; Munstermann *et al*., 2012).

The most challenging issue encountered was data quality issue of data granularity. This issue affected the studies that working with sensor data (Triki et al., 2015; Lltas *et al*., 2018; Wolf *et al*., 2013; Vitali & Pernici, 2015; Tax *et al*.,

2018; Munstermann *et al*., 2012). The second data limitation issue was inconsistent and incomplete data (Conca *et al*., 2018 & Najjar *et al*., 2018).

The applicability of process mining into the frail elderly domain was demonstrated in this systematic literature review which has just started gain attention in recent years given its small number of reviewed papers. On the other hand, no study was found that explore the management and care of frail elderly patient despite a growing number of elderly worldwide. It suggests extensions of work in applying process mining within frail elderly domain mainly involving the aspect of technical work. This reveals a crucial gap to fill in improving understanding the frailty and what is the best way to manage it clinically or non-clinically.

2.2.2 Process Mining Tools

There are two process mining tools used which authorize process mining techniques to create models, tables and data analysis; Disco (Günther & Rozinat, 2012) by flexion and ProM (Verbeek, Buijs & et al., 2010). The first tool used was Disco to explore the dataset and later with ProM for further analysis.

2.2.2.1 Disco by Fluxicon

Disco was a commercial tool which has been provided by Disco academic license and very user-friendly. This tool is used to work on the process analysis of raw data which create an automated process discovery, animation of the process map as well as detailed statistics and charts about activity logs of the dataset. It requires three specified parameters which are case Id, activity and timestamp from either MS Excel, CSV file, XES and XES.GZ files, MXML and MXML.GZ files, FXL and FXL disco log files. A visualise map of the process will be produced which mapping of each activity according to its particular timestamp will be shown after successfully importing the data. Furthermore, this process mining tool also provides filtering options; event log filter, timeframe filter, variation filter, performance filter, endpoint filter, attribute filter and follower filter. Each of these filters offers different capabilities of exploring onto specific case, event, timeframe and others to get further insight and analysis of the process.

2.2.2.2 ProM

ProM 6.6 is the main process mining tool implemented in Java that used to analyse the event log of MIMIC-III dataset (Veerbek, Buijs & et al., 2010). It is an

open source framework tool for process mining algorithms. Main supported import format file for ProM is XES (eXtensible Event Stream), and the others are MXML (Mining eXtensible Markup Language) and CSV files. XESame, ProMimport and Disco are tools used to extract the supported import format file for ProM. There are a variety of process mining techniques supported in ProM created in the form of plug-ins such as Heuristics Miner, Inductive visual Miner and etc. These plug-ins are available with the latest development in process mining with many different ways of filtering event logs to be applied to the data for process analysis.

2.2.3 Summary

Although both Disco and ProM authorize process mining techniques, there are several differences (Table 2.1) between them. The main difference is that ProM is an extendable framework which means researchers from academia and others are able to develop plug-in with advanced and improved process mining algorithms. In contrast to Disco, as the users are only able to use the available process mining algorithms provided in it. However, these plug-ins in ProM are usually complicated to use as compared to Disco.

Table 2. 1 Comparison of process mining tools features; Disco and ProM

Features	Disco	ProM
Easy to use	√	×
Open source	×	√
Extendable	×	√
Fuzzy Miner algorithm	√	√
Alpha algorithm	×	√
Heuristics algorithm	×	√
Inductive Visual Miner	×	√

Each tool offers great and unique ability to process mining techniques to be used on event logs. In this research and preliminary experiments, the first tool which is Disco was being used to obtain the initial discovery of the event log due to its easy to use the feature, and then ProM was being used later for further detail investigation.

2.3 Process Mining Methodologies

There is a number of methodologies followed for process mining to gain valuable information in any domain. In order to achieve that, a method needs to be developed in requiring successful completion of process mining application. In this section, there are two additional methodologies that support data mining project within the organisation. These methodologies are briefly described to show how process mining methodologies had been developed over the years following the available and established data mining project methodologies.

2.3.1 Crisp-DM

Cross-Industry Standard Process for Data Mining (Crisp-DM) provides as an outline to conduct data mining project and mainly focusing on the development and slightly into the management of the project. It consists of five phases; business understanding, data understanding, data preparation, modelling, evaluation and deployment (Chapman, Clinton & et al., 2000). During the first phase business project requirement need to be identified to ensure the knowledge discovery meet the requirement. Secondly, initial data will be obtained to identify any data quality issue and solve it to get it ready as the final dataset. Later in the modelling phase, various modelling technique will be applied using one of the tools; RapidMiner before being analysed and evaluated.

2.3.2 Crisp-TDM^n

Crisp-TDM^n is an extended framework from Crisp-DM where it focuses mainly on the temporal abstraction. The word is referring to the Temporal and n to multidimensional of clinical data (McGregor, Catley & et al., 2011). The extended task of crisp-TDM^n is at the first and second phase. In the first phase, two tasks was added; clinical objectives and data mining technique. Whereas in the second phase data description report was added with two additional attributes; data format and data quantity (Catley, Smith & et al., 2009). However, both Crisp-DM and Crisp-TDM^n were not created to fully support process mining methodology but specifically established methodology to support data mining project within organisations.

2.3.3 Process Diagnostic Method (PDM)

This methodology is a fast process diagnosis (Bozkaya, Gabriels & et al., 2009) to the whole overview of the process in a short amount of time. It consists of six phases; log preparation, log inspection, control flow analysis, performance analysis, role analysis and transferring result. Firstly, the extraction of event logs

from the information system is performed an an overview of the event log is required before analysis processes begin. During control flow analysis, information of process will be obtained based on the real execution of the process to look at whether it follows the recommended process flow. Later, questions regarding bottleneck of the process and the time taken to complete the process were done in the performance analysis phase. Whereas, in role analysis phase information is about who executes an event will be gathered by creating a role-activity matrix. Lastly, discussion with the domain expert will be held as to present all the result obtained and to conform the process model to the basis flow of their processes.

2.3.4 Rebuge's Methodology

This methodology is an extended work from the PDM methodology (Bozkaya, Gabriels & et al., 2009) as to deal with healthcare data characteristics which usually have variations in process and irregular behaviour. This work adds another step which is sequence clustering analysis after the log inspection phase (Rebuge & Ferreira, 2012). Clustering the log and pre-analyse the process was introduced to create a goal of the much simpler process model and schematize process analysis variations in process and irregular behaviour.

The steps are like a sub-process of the methodology where being executed in a cycle until the goal of creating a simple process model is achieved. The first step is to do sequence clustering in order to identify the behavioural pattern in the event log, next for cluster analysis to recognize which pattern symbolise the regular behaviour. Later, to understand regular behaviour and variants, and irregular behaviour. Before the sub-process ends, check if the process model is a simple enough for further cluster analysis, if not, it needs to undergo hierarchical clustering and start back from sequence analysis step again.

2.3.5 L * life-cycle Model

L* life-cycle model is a structured modelling approach for process mining project as shown in Figure 2.4. This life cycle model is defined in Process Mining Manifesto (van der Aalst, Adriansyah & et al., 2011) with six guiding principles to follow and prevent any mistakes done during process mining implementation. It consists of five stages; plan and justify, extract, create a control flow model and connect it to the event log, create an integrated process model, and provide operational support as shown in Figure 2.4 below. Plan and justification are required before starting the project, the type of data, objective and questions will be extracted. After getting all relevant data, control flow model is created which

is connected to the event logs at stage 2 and can be extended to other perspective during the next stage. Lastly, at stage four; intervene, adjust and redesign of the process would be made to achieve the objectives. The operational support stage is where any recommendation, prediction is proposed if only structured and stable process has been reached.

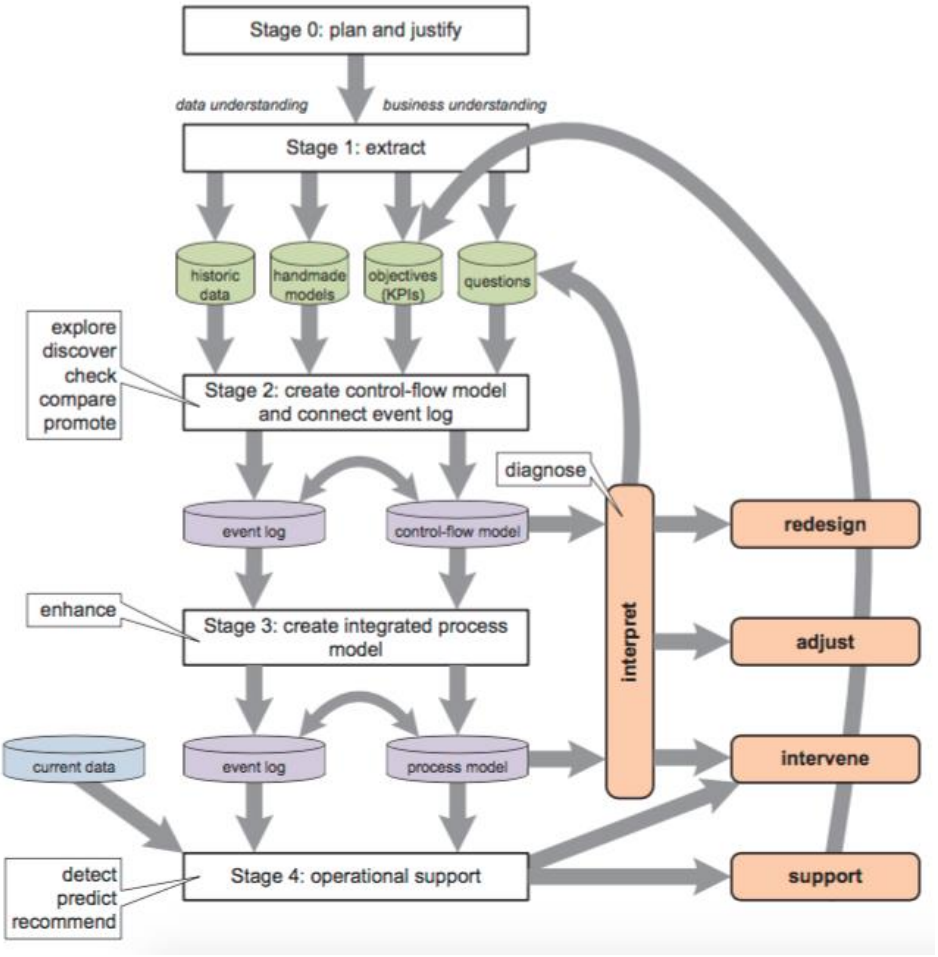


Figure 2. 4 An overview of L * life-cycle model in (van der Aalst, 2011)

2.3.6 PM2: Process Mining Project Methodology

Although, L* life-cycle model and PDM have been successfully adapted to mine process in healthcare project setting L* life-cycle model is not always suitable to be utilised for all kind of project. Hence, PM^2: process mining project methodology (Eck, Lu & et al., 2015) was introduced to support analysis of unstructured and structured process as well as covering many other process mining techniques. It comprises of six phases which are planning, extraction, data processing, mining and analysis, evaluation and process improvement and support. PM^2 will be the methodology of this study and explain in section 3.0.

The different type of input and output objects form the six phase of methodology that are briefly discussed as below:

- Goal related object
    - The project usually starts with project goals that creates research questions which will be fulfilled by performance findings and compliance findings which lead to improvements idea in order to reach certain project goal.
- Data related object
    - Data could be an input as well as output during the project and how it being manipulated to suit for certain phase. Firstly, data is extracted from any information system available in several forms, which associated to discrete event to form event data. Later, event data will be converted to event log that requires to have at least case id, timestamp and activity.
- Models related object
    - Two type of models could be derived are process model and analytics model. The first model could improve additional information provided by process such as resource consumption, data consumption and temporal restrictions. Whereas the second model would provide intuition into the process.

In the next paragraph, each phase PM^2 methodology will be discussed in detail:

1. **Planning**

   This phase is mainly associated with the goal related object whereby research questions, selection of business processes and creating project team are three actions that required to fulfil.

2. **Extraction**

   The objective and the output in this phase is to create event data mainly from the research questions and the available information system. Three actions that need to carry out are determining scope, extracting event data and transferring process knowledge.

3. **Data processing**

   In this phase, the input would be an event data obtained from last phase to create event log which will be used during mining and

analysis phase. Besides the event data, process models could also be used as the input. Creating views, aggregating events, enriching logs and filtering logs are the activities during this phase.

4. **Mining and analysis**

   During this phase, research questions will be answered in order to gain insight of process performance and compliance. Four types of actions in this phase are process discovery, conformance checking, enhancement and process analytics.

5. **Evaluation**

   After getting the result from the last phase, it will be evaluated to enhance the ideas for achieving the project goals. The actions that should be taken are diagnose, verify and validate.

6. **Process improvement and support**

   The evaluation result is the input in this phase to modify process for improvement. The main activities are implementing improvement and supporting operations.

2.4 Machine Learning

Machine learning approaches could be used to improve process mining in healthcare. Both machine learning techniques and process mining techniques are used for data analysis and to automatize knowledge acquisition. Process mining is a relatively new technique which is capable of analyse all end to end process of activities with the related resources. These activities that took placed in the process might be so complex to understand and machine learning technique could be used to identify the several distinct patterns of activities in process for analysis.

However, there are differences between process mining and machine learning. The main difference is the type of data used for analysis, process mining require data in the form of process related information data (event logs) (van der Aalst, 2011) whereas for machine learning only require information data (Alpaydin, 2010). Secondly, in order for machine learning to be able to classify or predict any pattern discover in the data, it needs a training datasets to learn the behaviour of the pattern in the dataset (Hastie, Tibshirani & et al., 2005).

Nevertheless, process mining do not have the ability to detect any pattern of process appear during the analysis. Process mining technique and machine learning technique will be discussed further in the next section of the report.

Machine learning is a data-driven method that iteratively learns from data to carry out specific task. Generally, it works by studying a set of data know as learning or training datasets in order to identify way of achieving goal. Then, it will be given new datasets that it never work on before to carry out the same task, and to validate its performance in accomplishing the task. This analytical model may be built predictive; to make prediction in the future or descriptive; to acquire hidden knowledge within the data or both (Alpaydin, 2010).

There are two types of learning strategies machine learning acquired; supervised and unsupervised learning. The only difference between these two learning strategies is that unsupervised learning does not have any output of the task from the data and need to define hidden structure within the data. Supervised learning includes classification and regression approach, whereas unsupervised learning includes clustering and dimension reduction approaches.

2.4.1 Machine Learning in Healthcare

Machine learning are commonly being applied with image data (Hassanpour & Langlotz, 2016), video and text. It is designed to discover statistical patterns in high-dimensional and multivariate datasets. There is significant work done in applying machine learning approach itself in healthcare due to its ability in analysing large set of data or big data.

The use of machine learning approaches in electronic health record whether to make prediction or to gain hidden information; descriptive are greatly depends on the intended situation of particular problem to solve. They are several predictive models of machine learning approaches applied in healthcare recently done that are being referred to as big group of statistical and mathematical methods that focused on prediction such as by Zhou, Gutierrez & et al., (2016) used classification technique to accurately predict a diagnosis of rheumatoid arthritis disease in secondary care, combination of three classifier; random tree, support vector machine and nearest neighbour using ensemble to preauthorisation decision by healthcare professionals (Araújo, Santana & et al., 2016), and developed a predictive model from baseline self-reports to predict the persistence and severity of major disorder (Kessler, Loo & et al., 2016).

2.4.2 Machine Learning in Process Mining

There are many works that has been done to improve process mining technique using machine learning. At first machine learning approach was applied to improve the quality of process model based on the four quality dimensions in process mining (Buijs, Dongen & van der Aalst, 2014) (van der Aalst & et al., 2012). The four quality dimensions of a process model obtained from event log are generalization, simplicity, fitness and precision. The authors also limit the search space of the algorithm by using process tree to overcome the previous genetic algorithm issues in (van der Aalst, Medeiros, & Weijters, 2005) such as deadlocks, improper termination and livelocks in process model.

Another process discovery improvement made is in supervised event abstraction problem of process mining to automatically transform low-level event log which is a complex and uninterpretable process model (also known as “spaghetti”-like model) to more high-level event log (known as “lasagna”-like model) (Tax, Haakma & et al., 2016) that aimed at improving the visualisation of process mining result. The study presented high-level traces prediction using evaluation metric that often applied time-window based method for sequence labelling field by generating a feature representation of the event log (XES log) with learning step from Conditional Random Field.

Furthermore, de Leoni & et al. (2016) made improvement not only for the control-flow perspective of process mining by proposed the clustering technique by correlating the independent and dependent characteristics of business process. The second study proposed improvement of the linear temporal logic (LTL) verification in process mining (Horita, Hirayam & et al., 2016). The LTL verification step done after the conformance checking completed. In this study, supervised machine learning technique was used to classify instances based on the predictor variables which is decision tree learning. This approach could effectively construct method of logical formulas to predict specific property each traces could hold concentrating on partial structures represented as event order relations of traces.

2.4.3 Trace Clustering

Besides that, many work also focused on finding the better understanding of process model as part of the discovery perspective of process mining. Song, Gunther & van der Aalst (2009) used trace clustering to cluster different processes from process log per case similarities of profiles to improve process

model discovery accuracy. They used several distance measures (i.e. Euclidean distance, Hamming distance, Jaccard distance) to calculate similarities between cases before applying clustering technique such as K-means clustering, Quality Threshold Clustering, Agglomerative Hierarchical Clustering and Self-Organizing Map which implement divide-and-conquer approach. From case study using hospital record of gynaecological oncology patient, combination of Euclidean distance and SOM performing better than other that include case and event attributes in trace profiles.

Later, Song, Yang & et al. (2013) enhance the trace clustering performance by incorporating dimensionality reduction. The experimental results showed increase in computational time spent on the subsequent clustering task, and in some cases improved result of the cluster formed, depending on the combination of techniques used and the context of the dataset. For artificial neural network self-organizing map (ANN SOM), clustering result improved only with simple dataset, regardless of the dimensional reduction techniques used such as singular value decomposition, random projection and principal component analysis. Although not discussed by the authors, the fact that ANN SOM did not benefit from the prior dimensional reduction as much as k-means, ANN SOM could make projections of high-dimensional spaces for low-dimensionality spaces that make it easier to analyse data in higher dimensions.

Markov cluster algorithm was applied in process mining technique as a plug-in in ProM by Hompes, Buijs & et al. (2015) to independently discover number of clusters and group cases of similar behaviour based on specific perspective. The behaviour of process discovered could be normal or deviate behaviour. In the same year, Hompes & et al, extended the work done previously by providing new technique in comparing clustering which is change point in behavioural similarities between cases. The authors used trace clustering to find common and deviating process behaviour by looking both at control flow and data attributes that make the technique as context-aware.

2.4.4 Time Prediction

Furthermore, implementation of machine learning in process mining used also to predict the completion time of running instances (van der Aalst, Schonenberg & Song, 2011) by extending the discovered process model. They proposed a transition system to predict completion time of an activity in a complete process and developed a plug-in in ProM called FSM Analyzer which takes transition system from FSM Miner (Aalst, Rubin & et al., 2010) and an event log as input.

Later, Polato, Sperduti & et al. (2014) improve the transition system by implementing classification and regression models. The transition system had three extensions, where the first one encodes average time spent in every nodes. In order to identify the probability distribution over the states reachable from the current, it used Naives Beyes classifiers in every states. Lastly, support vector regressor will predict the completion time at each transition.

Chapter 3

Research Methodology

The previous chapter has demonstrated the background on understanding frailty domain in elderly and technicial aspect of process mining and machine learning. This chapter discusses the methodology used throughout this work to execute the analysis implementing process mining and machine learning techniques. It includes details of each stages in the methodology and the extended approaches adopted. The methodology of this study is established on the characteristics of healthcare dataset which is complex and the appropriateness of existing techniques toward the frailty progression analysis. The methodology in this study has been accustomed towards achieving the goal of each experiments in frailty progression analysis.

3.1 Methodology Expansion Strategy

The research methodology established in this work was from two pioneer process mining projects methodology: (1) L* life cycle methodology (Van der Aalst *et al.*, 2012), (2) Process Mining Project Methodology (PM2) (Van Eck *et al.*, 2015) and and a healthcare specific methodology known as Process Mining Methodology for Exploring Disease-specific Care Processes (MEDCP) (Vathy-Fogarassy, Vassányi and Kósa, 2022). Table 3.1 illustrates the summary of the expansion strategy to generate the methodology for this work and its main stages.

The referred methodologies and this work methodology start by ***Stage I*** (*Planning*) or Stage 0 in L* life-cycle. It comprises of understanding the domain and the available dataset used in the project, in L* life-cycle methodology it consists of selecting business processes, identify research questions and finding project team, while in MEDCP it composes of finding, collecting and integrating the raw data from different sources of EHR database. In this work methodology, Stage I includes generating aim, scope and developing the research questions. The component Stage I of this work methodology are the extended step after understanding the domain and dataset (from L* life-cycle), choosing the business process (from PM2) and after obtaining the dataset as explained in steps of methodology (MEDCP).

Table 3. 1 The summary of expansion strategy of the work methodology

Methodologies for Process Mining Analysis			
L* Life-Cycle	**PM2**	**MEDCP**	**This work Methodology**
0. Plan & justify	1. Planning	Find raw data	1. Planning
1. Extract	2. Extraction	Migrate & transform data into event log	2. Extraction
Not specified	3. Data processing	Preprocess data	3. Data Transformation and Loading
		Create multi-level abstraction	
2. Create control-flow and connect event log	4. Mining & analysis	Domain-specific framework	4. Mining & analysis
3. Create integrated process model		Process mining tool	
Interprete (from the side of methodology)	5. Evaluation	*Not specified*	5. Evaluation
4. Operational support	6. Process improvement & support		*Not relevant*

Stage II (*Extraction*) is the second stage following the planning stage. In this work methodology the extraction stage was expanded to include developing the selection and exclusion criteria required to extract the study cohort from the dataset population. It was done after transferring the process knowledge (from PM2 activity) which require the researcher to understand the background of the process knowledge beforehand. The migration and linkage work (activities in step two from MEDCP methodology) has been achieved in this work methodology as both datasets have been integrated into a single database management system and direct linkage of all care events was done through the database. The study cohort was extracted after the data quality inspection (explained in Section 4.2.5 for MIMIC-III dataset and Section 5.5 for Bradford dataset) was done following the data quality element in (Kahn *et al.*, 2016).

Stage III (*Data Transform and Loading*) of this work methodology comprises of combination steps of processing and transformation of raw dataset (study cohort) into event log required for process mining analysis. It combines the components in the data processing in PM2 methodology, and the proprocess data and creating multi-level abstraction as explained in MEDCP methodology. However, this step is not specified in detail for the L* life-cycle methodology. The data transformation consists of log filtering, events abstraction, determining the frailty score, log enrichment, securing the event sequences and aggregating events that will be discussed in further.

The next stage is ***Stage IV*** (*Mining & analysis*) similar as the PM2 stage which is the crucial part of the methodology. The L* life-cycle methodology set up the

guidance for mining and analysis in the creating the contro-flow model and connect event log. It consist of process discovery analysis and conformance checking. While in the create integrated process model composes of the enhancement step. The methodology in this work, follows both the PM2 and MEDCP methodology. It combines the process discovery (using any existing process mining tools as in MEDCP methodology), process analytics, conformance checking (as list out in PM2), and developing domain-specific framewok employed in MEDCP for frailty progression analysis. In this study, the frailty progression analysis comprises of utlising the process cube-based analysis, determining frailty trajectories as part of frailty progression and confirmatory analysis. The confirmatory analysis is done adopting the discretization method before conducting the comparative analysis in identifying the difference between pattern of progression within two logs.

The final stage of this work methodology is ***Stage IV*** (*Evaluation*). The evaluation part in the L* life-cycle methodology was not specifically included in the main stage. It consists of interprete, redesign, adjust, intervene and support the analysis. In the PM2 the steps include diagnose, verify and validate. While in MEDCP no evaluation step is specified. In this work methodology, analytical assessment, domain expert input and confirming with the literature are the steps taken in this stage.

The operational support (in the L* life-cycle) and process improvement and support (in PM2) are not included in this work methodology. Both components recommended the implementation of the real-life analysis finding (as part of improvement purposes) into the process. However, this steps are not included in this work methodology as it isout of the research scope.

3.2 The Methodology

The methodology used in this work was built based on the expansion strategies discussed in the above section. It is established based on the PM2 and the MEDCP methodology and further expand to make it relevant for frailty progression analysis. The combination or expansion from the PM2 and MEDCP were done as to improve the methodology specifically for process mining project in disease-based analysis. The stages of the methodology are: Stage I (Planning), Stage II (Extraction), Stage III (Data Transformation and Loading), Stage IV (Mining & Analysis) and Stage V (Evaluation) will be discussed further in the next sub-section.

3.2.1 Stage I: Planning

This stage intention is to develop aim, research questions, proper team and limitation as the initial point in process mining project. The development of the components required in Stage I was following the PM2. The first component is selecting the process for analysis, identifying research questions, aim and score, and lastly is composing team project.

The first component in this work is selecting the process to be analysed. Frail elderly pathway with the focus on frailty progression is the process to analyse in this work. The study population selected is elderly population. The next component is identifying research questions, aim and limitation of the project. This work followed the combination of question driven and data driven process mining project to answer specific questions regarding the frailty progression as the process. Some questions include (i) what is the most followed path in the elderly population? (ii) Is there any activities that speed up the frailty progression? and (iii) Are there anny difference in frailty progression within different frail elderly group? are from the question-driven project. While data-driven project questions derived from the initial data analysis, such as confirmatory analysis. The last component of this stage is composing a project team which was identified at the early stage of the project. In this work, the team of which the researcher are working together is from Bradford Institute of Health Research (BIHR). The team comprises of the leading expert in frailty is Professor Andrew Clegg and database management member. The domain expert is necessary in reviewing the project and database management team in assisting the EHR of study population.

3.2.2 Stage II: Extraction

The second stage involves the extraction of data from both MIMIC-III and Bradford dataset. The direct access of the MIMIC-III was achieved by downloading the full set of dataset from the openly available healthcare website, Physionet and re-build the database in the PostgreSQL database management system. Whereas for Bradford dataset, the access was only done using the facility provided from BIHR. The extraction of the study cohort done following the planning agreed on the previous stage and based on the focus of each analysis. The involvement of the team members of the project are essential in understanding the nature of the dataset with regards to how the data was generated into EHR, the specificity of the dataset and attribute required for analysis.

3.2.3 Stage III: Data Transformation and Loading

In this stage, event logs will be created as the output. The data transformation stage is following the PM2 processing stage. It aims to transform the raw study cohort extracted in previous stage into event log for process mining analysis. It include log filtering, event abstraction, determining frailty score, log enrichment, securing the event sequence and aggregating event. The transformation steps are applied in the extracted cohort according to the process mining project specification. Each of the data transformation step is discussed further.

3.2.3.1 Determining the Frailty Scores

Frailty scores are the indicator used to stratify the elderly patient as non frail or frail, with frail elderly is further grouped based on the severity of frailty following the approach in (Clegg *et al.*, 2016). The score was determined based on the total number of accumulated deficit associated with frailty a patient acquired at certain point within the study duration. The deficits are defined as age-related health issues that are prevalent with elderly which include diseases and disability. The list of age-related health issues associated with frailty deficits followed by Clegg et al., (2016) with a total of thirty-six as shown in Chapter 5 Section 5.8.1.2 in Table 5.7. Each of the frailty deficits comprises groups of procedures, clinical and non-clinical findings known as clinical coding, Read Code as further explains in Chapter 5 Section 5.3.1. The accumulated unique deficits acquired by a patient over time is divided by thirty-six to get the frailty score at that point of time. Table 3.2 presents the range of accumulated deficits and score range in determining frailty category of the elderly. The identification of frailty scores was done at each visit to the GP practice.

Table 3. 2 List of range of frailty score and category

Category	Accumulated Deficit Range	Score Range
Clinically Fit (Non frail)	1 – 4	0.03 – 0.11
Mild	5 – 8	0.14 – 0.22
Moderate	9 – 12	0.25 – 0.33
Severe	13 – 36	0.36 – 1.00

3.2.3.2 Event Abstraction

The simplest events abstraction is event aggregation which involves events with characteristics of low-level granularity. It aims in reducing the complexity to enhance the comprehensibility of process model. The event abstraction is done by modifying the level of details from the low-level events from the rescource or

activity attributes. An ontological concept facilitates the events abstraction transformation step and validation from the clinical domain is required. In this work, two events abstraction were conducted on the resources events in Section 5.8.1.1 and on activity events in Section 5.8.2.1.

3.2.3.3 Log Enrichment

Although creating the simplest form of process model is vital in process mining analysis, adding additional events into the log is necessary in having comprehensive model to analyse. The log enriching step is following the PM2 method deriving from computed events as in this work. The computed frailty scores were transform and added into the log as an event of frailty category. This is to mark the transition between frailty stages from fit to mild, mild to moderate and from moderate to severe.

3.2.3.4 Securing the Event Sequences

Securing the events sequence was done to fix the temporal inconsistency issue after adding additional events into the log. Two procedures are taken in performing this data transformation step: (1) sorting the events with the same date (timestamp) alphabetically based on the activity name and (2) re-arrange the frailty category events with the same date as deficits events. Example for re-arrange event is Fit event is always come first before the deficits events, while the other frailty category events goes after the deficits events. The justification and example of this transformation step are discussed in Section 6.4.3.

3.2.3.5 Log Filtering

Log filtering is a common data transformation step. It aims in producing a simple and understandable process model. The log filtering is based on PM2 method which includes variant based filtering and compliance based filtering. The first log filtering is to select the similar traces as conducted in experiment 5 where the less dominat traces was filtered out from analysis. Whereas, the second log filtering is compliance based filtering is applying a set of rules in ensuring the traces that fit or suitabile to be included in process model. For example in this study, elderly with no frailty deficit is excluded in investigating the frailty progression.

After data transformation steps were applied, an event log is created in the form of .CSV file or .XES file as part of the Loading stage. The .CSV file is directly created after the transformation step done using the Python Notebook Jupyter

application. Later, the .CSV file is loaded into the Disco tool to generate .XES file for analysis in ProM tool.

3.2.4 Stage IV: Mining & Analysis

Process analytic and discovery are the main activities in this phase. The mining analysis are performed using both Disco and ProM tools. Disco tool was used for a quick analysis to gain insight on the event log and process model with the statistical analysis automatically generated. Disco is usually the preferred initial analysis tool as it provides the best user friendly interface tool. While, ProM is used to perform advanced analysis on the event logs and process models. Furthermore, other statistical analysis was done using Tableau tool to visualise or identifying the significant behaviour in the datasets. Two techqniues of process mining are included in this work (1) process discovery and (2) process conformance. Process analytics obtained through the systematic computation analysis of the process or event log. It can either be automatically generated from any process mining tools or computed through the Notebook Jupyter.

3.2.4.1 Process Discovery

The input of the process discovery is an event log to produce a process model without using any insights in advance. Process discovery technique is the essential part in process mining analysis. It is a sequence of activities illustrated in a network like model or graph. In this work, the common process model used to visualise the process directly-followed model and transition system model.

3.2.4.1.1 Directly-Followed Model

The directly-followed model is commonly used process model in this work. It is can be generated using Disco following the fuzzy miner algorithm or ProM using the interactive Data-aware Heuristics Miner (iDHM) plug-in (Mannhardt, De Leoni and Reijers, 2017). The process model from Disco was used because it offers user-friendly interface and stable during the execution process. Whereas, the directly-followed produced by iDHM plug-in was used for smooth integration flow from process discovery to conformance checking of process model in ProM.

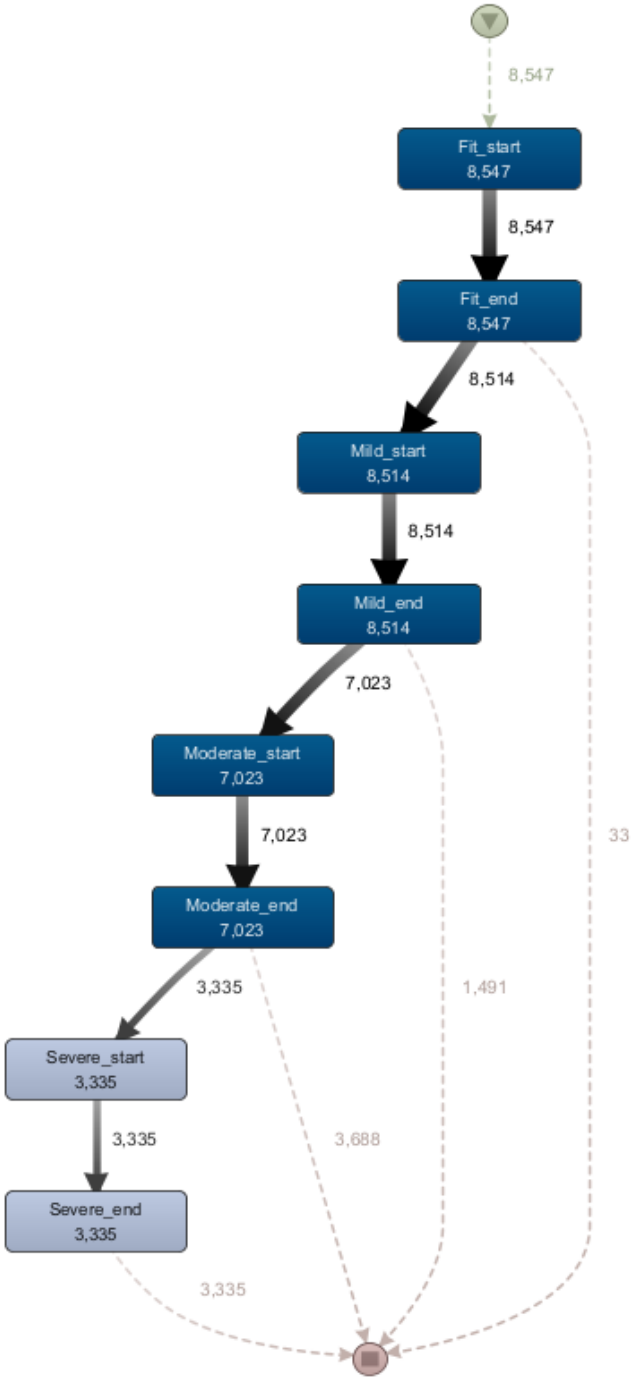


Figure 3. 1 A directly-followed model from Experiment 6 in Section 6.4.4 using the frequency perspective view of process model

Figure 3.1. illustrates the directly-followed model produced by Disco. The flow of the model in Disco usually represented from top to bottom with the starting point is the green circle at the top of model and end with a red square at the bottom of model. The box represents the activity of the process with the egdes connecting the boxes are the flow of one activity to the next activity. The different shades of boxes and edges reporesents the frequent executed activities or flow of activity. The darker shade indicate more frequent activity and flow. The input of any process mining tool is an event log in the file format of .CSV or .XES. The .XES file can be transformed in the Disco (generate .XES file from the .CSV file) and ProM (using the plug-in “Convert CSV to XES”).

3.2.4.1.2 Business Process Tree

The other commonly used process representation for analysis is business process tree. The model can be generated in ProM tool using the Inductive Miner (IvM) plug-in (Leemans, Fahland and Van Der Aalst, 2014; Leemans, 2017). Figure 3.2 illustrated the process tree using the default setting implemented in Experiment 1 in Section 4.2.4.

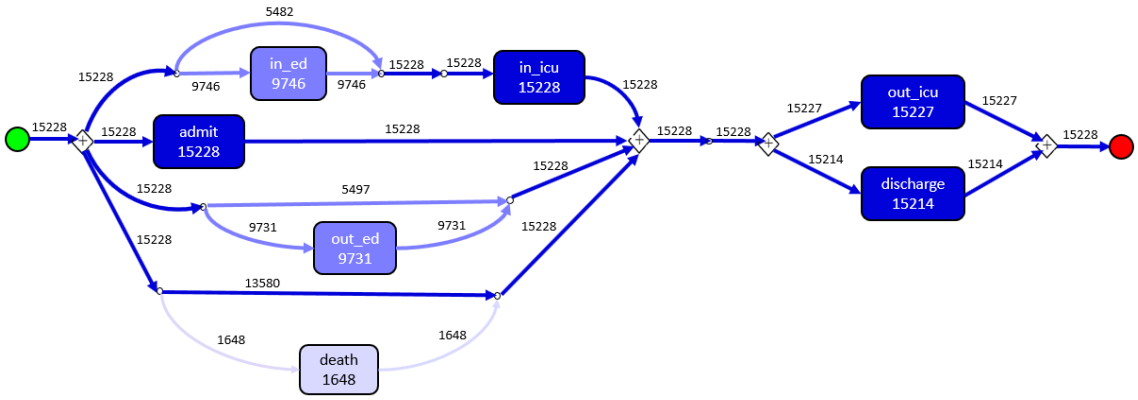


Figure 3. 2 Process tree produced from the IvM plug-in in ProM. Implemented in Experiment 1 in Section 4.2.4

The model shows the flow of process in a sequential manner. The rectangle shapes represent the activities in the process with darker shades rectangle indicates the higher frequency of activities executed compared to the bright shade rectangle. The nodes with symbols express the behavior of the model. The green circle represents the source of the process while the red circle indicates the end of the process. The two nodes illustrated in Figure 3.2 of ‘+’ in diamond shapes is concurrency denotes that all activities need to be executed even at the same

time and the small circle nodes represents the exclusive choice where only one of its children need to be executed. Meanwhile, there are two other nodes not illustrated in the model is interleaving node with symbol '↔' in diamond shape that means all activities need to be executed separately. The node with symbol 'O' in diamond shape indicates inclusive choice where at least one of the children need to be executed while multiple children is allowed to execute at the same time.

3.2.4.1.3 Transition System Model

The second process model is transition system model produced using the Process Comparator plug-in in ProM (Bolt, de Leoni and van der Aalst, 2018). The implementationof this plug-in are presented in the Experiment 5 in Section 6.4.4.2 and Experiment 7 in Section 7.4.2.4.

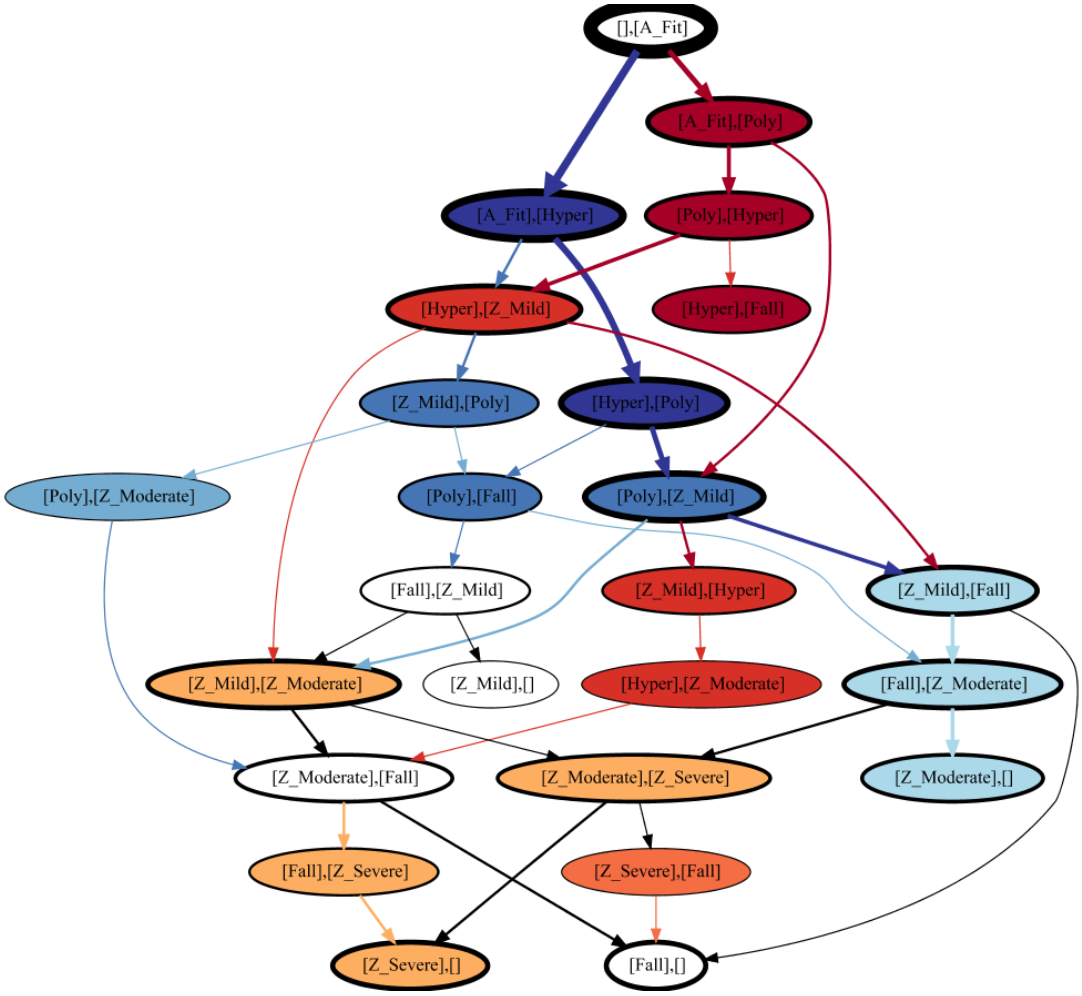


Figure 3. 3 A transition system model produced using the Process Comparator plug-in utlising the second type abstraction. The abstraction illustrates the model by combining the direct sequence of the last two activities into a node of actvitity in transition system model.

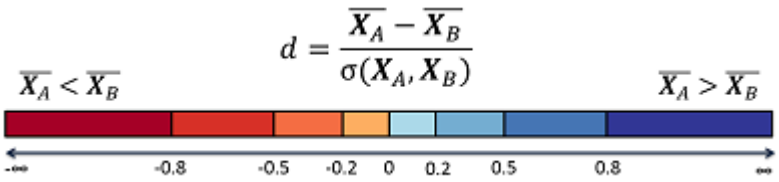


Figure 3. 4 The color legend of process comparator plug-in.

Figure 3.3 shows the transition system model. The aim of the plug-in is to determine the differences between two event logs represented using the color legend Group A and B as shown in Figure 3.4. The shades and thickness of the nodes and edges represents the behavioural properties of the process such as frequency, duration or elapsed time (can be adjusted in the plug-in comparison settings) executed between two nodes or transition between nodes. The shdes of red color represent Group B with darker shade indicate the metric is larger in Group B, while, the blue shades color represent the higher metric in Group A as presented in Figure 3.4. The white nodes represents the statistically significance different between the two event logs is not exist. The statistical significant different test is following the standard alpha level of 0.05 and is changeable in the plug-in setting.

3.2.4.2 Other Visual Representation for Pathway or Process

There are two other visual representation that heavily used to facilitate the analysis in this work. The first is dotted chart and trace variant diagram which will be discussed in the following section.

3.2.4.2.1 Dotted Chart

A dotted chart is a visual representation of the activities distribution from each traces in the event log across time. It can be produced using the visualisation option of dotted chart (Log projection) in ProM. Figure 3.5 shows the an example of dotted chart from Section 5.4.2. The activities are represented as the colourful dotted in the diagram. The x-axis indicated the duration of traces in projection from start to the end and y-axis is the traces of the patient. The diagram has been sorted from the longest traces to the shortest.

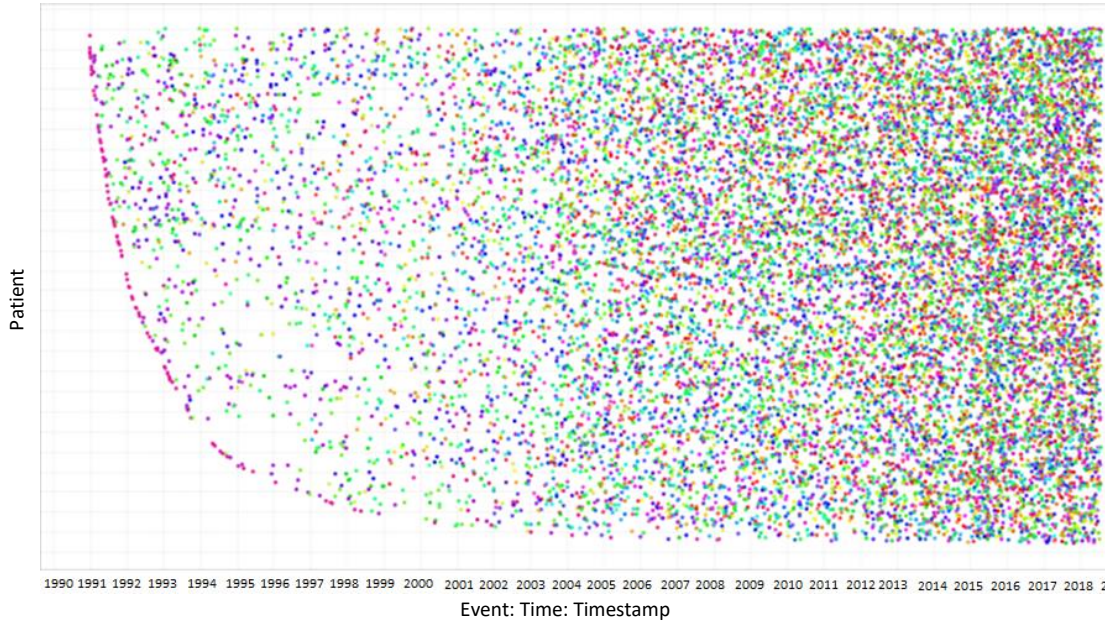


Figure 3. 5 The color legend of process comparator plug-in.

3.2.4.2.2 Trace Variant Diagram

The second additional visual representation is trace variant diagram generated from the ProM. It illustrates the sequences of the dominant or frequent sequence of activities as trace variants. It can be produced using Disco and ProM. Figure 3.6 shows an example of trace variant diagram from Experiment 1 in Section 4.2. The diagram is generated in Disco through the cases tab features where it display all variants in an event log in ascending order. While in ProM, the diagram is generated through the visualisation option listed as Explore Event Log (Trace Variant/Searchable/Sortable) (Enhancemenet).

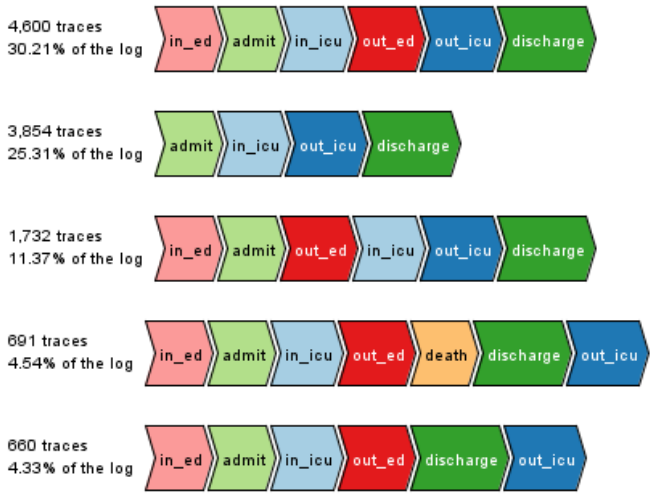


Figure 3. 6 The trace variant diagram generated in ProM

3.2.4.3 Conformance Checking

The last process mining technique implemented in this work is conformance checking. The aim is to check whether the reality (what happen in the model) conform with the recorded log. The conformance checking determine the three quality checking metrics of process model namely fitness, precision and generalisation. The commonly used plug-in to identify these three metrics are (1) Replay log on Petri Net for Conformance Checking and (2) Measure Precision/Generalisation.

The first plug-in needs two input files to execute the coformance checking between the model and the log. It requires an event log and a model in the notation of Petri Net. It identify the mismatch value (fitness value) between the model and log by measuring the skipeed activites (activities should happen in the model but not exist in the log) and inserted activities (activities were not observed in the model but appear in the log). The fitness value can be obtained through the inspector display of global statistics as shown Figure 3.7. Furthermore, the details of which the mismatch occurred can be observed through the visualisation options available, Project Alignment to Log (PNetReplayer) in ProM.

Figure 3. 7 The Trace Fitness value using the Replay log on Petri Net for Conformance Analysis plug-in in ProM

The second conformance checking metrics are precision/generalisation which can be generated using plug-in Measure Precision/Generalisation in ProM. It requires three input files (1) the event log, (2) the Petri Net and (3) the net replay result obtained from the previous Replay log on Petri Net for Conformance Checking. The result from the plug-in Measure Precision/Generalisation will produce the two metric values.

3.2.5 Stage V: Evaluation

The aim of the evaluation stage is to determine whether the result of analysis is useful and whether the developed appraoach could answer the research questions. It requires the the opinion and knowledge from the domain experts through series of discussions. The evaluation is determined through the analytical assessment obtained from the mining & analysis result. In addition to that, statistical assessment was conducted as part of evaluation stage. It is evaluated using the strength association of Relative Risk (RR) score, binomial test, and correlation coefficient test. The statistical test will be discussed in the next section. Furthermore, the evaluation also was done through the confirming steps between the finding of analysis with the finding reported in the literature or past studies. The aim is to observe whether the finding obtained in this work is similar or difference from the reported results.

3.3 Frailty Progression Analysis

The frailty progression analysis is a time-based analysis. The main task are the performance assessment and combination of questions and data driven appraoch. The analysis is dependent on the frailty identification tool developed by Clegg and team (frequently mentioned as Clegg's approach in this study). The frailty identification tool is useful in measuring the transition of frailty stage or state from low to high severity level. Frailty is analysed based on two different task (i) the trajectories and (ii) the progression via the significant pattern.

3.3.1 Trajectories and Significant Pattern of Frailty

The frailty trajectories analysis is following the approach of disease trajectories in this past studies (Jensen *et al.*, 2014). As the graph or network like model was produced was produced without utilising the process mining technique, it was formed based on the concatenation method of frequent and significant trajectories of bi-directional pair of disease. However, recently the work have

been successfully applied using process mining to examine the disease trajectories (Kusuma *et al.*, 2020, 2021).

3.3.1.1 Quantifying Disease Association

The steps in investigating the frailty trajectories is starts with filtering out the directional adjacent pair of frailty deficits with less than 10 occurrence. It is to ensure only the dominant directional adjacent pair within the study population is selected for the analysis. Next, disease association or frailty deficits association was quantifed within its adjacent deficits. The diagnosis association measurement implemented is the Relative Risk (RR). The RR is the ratio of probability of an event occurring in the exposed group versus the probability of the event occurring in the nonexposed group or also known as risk ratio is risk comparison of health event between exposed and unexposed group (Morris and Gardner, 1988). It provides an information of a likelihood of the event occurring in comparison between the exposure and non-exposure group.

Each pair of diagnosis-based deficit strength of association is determined using the Relative Risk (RR) score. It defined as risk of developing a subsequent disease after having the first disease is related to the risk of having the first disease and the subsequent disease due to the baseline of occurrence. The equation to measure *RR* is given in the equation as:

$$RR_{ij} = \frac{N_{ij}/[N_{ij} + (N_0 - N)]}{N_{ij}/[(N_j - C_{ij}) + N_0]} \quad (1)$$

The patients who diagnosed with diagnosis deficit of *i* and *j* are represented as N_i and N_j, while the combination of both diseases a patient had denoted by N_{ij}. Meanwhile, C_{ij} represents the combination of diseases *i* and *j* in order of *i* first then *j*. Lastly N_0 represents number of patient unaffected with either of diagnoses. The *RR* value of 1 indicates that the risk of having health event is similar between exposed and unexposed group. Meanwhile *RR* value less than 1 means the risk is lower in exposed group and if the *RR* more than 1 indicates the risk is increased in exposed group.

A second filter is applied by calculating the Confidence Interval (*CI*) which aims to provide the precision of the result. It give range of values of which true value is exists with certain confidence level (Tseng and Flechner, 2011). The second

filter is applied by calculating the 95% of *CI* of *RR* with the following given equation following the (Morris and Gardner, 1988):

$$CI = \exp\left(\log RR \pm \left(N_{1-\frac{\alpha}{2}} * SE(logRR)\right)\right) \quad (2)$$

The standard error of log *RR* is defined as:

$$SE(\log RR) = \sqrt{\left(\frac{1}{N_{ij}} + \frac{1}{N_j - C_{ij}} - \frac{1}{N_{ij} + (N_i - C_{ij})} - \frac{1}{N_j - C_{ij} + N_0}\right)} \quad (3)$$

The *CI* is calculated separately for each frailty categories. Only selected pair with *RR* value more than 1 and *CI* value is statistically significant if the value 1.01 is included if it out of the range of confidence interval. The deficit diagnosis co-occurrence of which increase the risk of subsequent event by more than 1% is the focus in this experiment.

3.3.1.1.1 Test of directionality

The evaluation of directionality of pair of deficit diagnosis is performed using the Binomial test. The deficit diagnosis pairs of (*i*, *j*) which had *RR* > 1 and *CI* value 1.01 out of the range of both directions ($D_i > D_j$) and ($D_j > D_i$) are tested for directionality. The directionality test of pair of deficits is to determine which direction of ($D_i > D_j$) or ($D_j > D_i$) is having significant direction within the cohort with the *P*-value is less than 0.05. The test of directionality is evaluated using the binomial test from the Python `scipy.stats.binomtest` package.

3.3.1.2 Event Log Creation

The event log for process mining analysis is created using the pairs of deficits diagnosis with their significant direction. In this stage, each patient diagnosis log is assessed to only contain pair of deficit diagnosis that is significant and the co-occurrence of the exposure increased risk of having the subsequent diagnosis. The recurring diagnosis is removed for each patient and only first occurrence is kept as the final processing step.

The event log comprises the significant trajectories of frailty deficits is then loaded into any process mining tools for analysis. The investigation of frailty trajectories using Bradford dataset is presented in Experiment 3 in Section 6.2. THe frailty

trajectories in Experiment 3 was presented using the plug-in Directly Followed Visual Miner (DFvM) in ProM.

3.3.1.3 Quantifying the Rate of Frailty Progression

The electronic frailty index (eFI) score used as one of the frailty measurements tools as a conceptualisation of frailty. It is based on the accumulation of the deficiencies across multiple bodily systems including physical, cognitive, and social. Although eFI useful in quantifying the level of frailty in a person, additional information such as speed of frailty progression could be beneficial. The speed of frailty progression over time is calculated following the formula:

$$Rate\ of\ progression = \frac{\frac{y_2}{36} - \frac{y_1}{36}}{t_2 - t_1} \quad (4)$$

The two consecutive visits to GP were identified to measure the rate of frailty progression also defined as progression point illustrated in formula (4). The eFI scores at both visits were determined to find the value of eFI score increment, where y_1 and y_2 are total accumulated deficits at visit 1 and 2 respectively. The interval between visits 1 and 2 were computed and indicated by t_1, time at visit 1 and t_2 time at visit 2. The progression point between two visits are calculated by dividing the eFI score increment with the time interval between two visits.

3.4 Confirmatory Analysis

The confirmatory analysis is consists of an approach in determining whether the cut-off points developed in Clegg's appraoch is difference or similar with the proposed cut-off points of this work. The comparison of the two approaches is done through comparative analysis using the plug-in Process Comparator in ProM. The discretization is implemented in determining the best cut-off points utilising the target features to facilitate the optimal splitting points.

3.4.1 Discretization Method

Discretization is an important pre-processing step of data reduction approach in machine learning, data mining task and knowledge discovery. It involves reducing or transforming the data distribution to finite range of interval where each range represent several categories (Dash, Paramguru and Dash, 2011). Discretization normally performed ahead of the learning procedure to create a concise abstract

of continuous attributes. It helps in speeding up the learning procedure accurately and facilitating the end users and experts in understanding the data (Liu *et al.*, 2002).

Several dimensions of discretization approach extensively proposed in the literature combining different dimensions to create a unique method. Binning is a discretization approach usually employed (at least one and not only restricted to two dimensions): splitting or merging, incremental or direct and either unsupervised or supervised (Liu *et al.*, 2002). Unsupervised binning discretization is the simplest approach which offer two strategies of binning by either equal frequency or equal width.

There are two approaches in performing discretization in this work. Both type of discretization is part of data-driven machine learning technique.

3.4.1.1 Discretization based on *K*-Means Clustering

A high volume of progression points identified from previous step is further transformed to determine the type of frailty progression. It was evaluated to define the progression point as progressing drastically or high, medium progress, and slowly progress. Hence, a robust way of splitting points to distinguish different type of frailty progression was done using the discretisation approach. Discretisation is a conversion of a continuous variable into discrete attribute, and it is widely used in sociological data analysis to aid the comprehension by combining numerous values of a continuous characteristic and separating the continuous domain into non-coinciding intervals. An example of application of discretisation is the transformation of numerical value of human heights into discrete value of short, medium, and tall.

In this experiment, discretization based on *k*-means clustering method is applied on the continuous values. The approach is chosen over the two simplest discretisation methods as they possess limitations such as producing an uneven number of data points at some intervals and an overlapping problem as same occurrence may resides in difference interval bins (Dash, R., Paramguru and Dash, 2011). The two methods are equal-width interval discretisation and equal-frequency interval discretisation which are based on the equal width or frequency of interval binning.

K-means discretisation approach is an unsupervised clustering technique applied on the 1-dimensional continuous input values. It will find the optimum number of intervals as the splitting points to define the frailty progression. The approach follows distance-based similarity measures and calculated to cluster the date

points with defined number of clusters, *k* (Palaniappan and Hong, 2008). The aim of clustering is to seek similar instances and group them into clusters such that gap between instances inside the cluster is smallest as possible and distance between clusters is largest as possible. The formula (5) is the approach taken to split the continuous data points by optimizing the objective function following Xu He, Fan Min and Zhu, (2014):

$$H = \sum_{i=1}^{k} \sum_{x \in C_i} d\,(x, a_i) \tag{5}$$

The formula (5) denoted C_i as the centre of the cluster stated by a_i. The Euclidean distance is being measured between the data points x and a_i, indicated as d (x, a_i). The shortest distance between the respective clusters centre and data points is the aim of the objective function, H stated in (5). Initiation of splitting points also known as clusters centre are done randomly by the algorithm based on number of k. The initial data point distribution was determined by assigning data points to the closest cluster centre. Next, new clusters are formed by calculating the former cluster centre as average of all data points within each cluster. The step was done iteratively until a condition called convergence is met. The convergence happened when the sum of squared distances are minimised between different cluster over data points in each clusters (Gupta, Mehrotra and Mohan, 2010).

3.4.1.2 Optimal Binning Discretization

Supervised binning process able to reduce data distribution complexity through data transformation. It is beneficial in interpreting the arbitrary reliance between feature and target variables. Binning process implementing machine learning classifier of decision tree algorithm that measures the preliminary split data points to create pre-bins (Navas-Palencia, 2020). It composes of two steps, 1) preliminary granular discretization is created during pre-bin process and 2) an iterative optimisation to fulfil certain constraints. The iterative step is a merging of consecutive pre-bin computed by the discriminant level identified as information value to find the optimal split points.

The approach requires only single target variable, hence data dimensionality technique is implemented. Principle Component Analysis (PCA) is an unsupervised machine learning approach that linearly transform the dataset to preserve its maximal variance (Jolliffe and Cadima, 2016). A standardisation of target variables is performed earlier using the Scikit-learn function of *StandardScaler* (Pedregosa *et al.*, 2011). The principal components are

evaluated with a formula of $StandardScaler\ value\ of\ x_i = \frac{x_i - mean\ of\ x}{standard\ deviation\ of\ x}$ where x is the target variable. The purpose is to normalise the range of variance between target variables. The PCA is sensitive to variances, hence standardisation prevents the dominating of larger scales of target variable over the smaller scales. The optimal binning approach is implemented to discretize feature variable based on target variable obtained through the PCA step. The function `optbinning` for continuous target from python library OptBinning (Navas-Palencia, 2020). The number of bin is set to four as represented in Experiment 6 in Section 7.3.4 similar to the number of frailty categories reported in past study (Clegg *et al.*, 2016).

3.4.2 Correlation Coefficient

A statistical method is used to calculate the association between variables. Correlation is statistical tool to assess the level of association between two quantitative variables of each groups. Two types of correlation coefficients methods are widely used: Pearson's correlation coefficient and Spearman's rank correlation coefficient. The assumptions of Spearman's Rank are required to be type of ordinal and not limited to continuous variables. Apart from that, it evaluates the association between variables with monotonic relationship (Schober, Boer and Schwarte, 2018). It describes as direct or indirect relationship between two variables which both variables increase at different rate or one variable increase and the other decrease.

3.4.2.1 Spearman's Rank Correlation Test

Spearman's rank correlation is non-parametric test where Gaussian distribution of data is not assumed. It is best to apply when data distribution is skewed or ordinal in one or both variables. The formula to calculate the correlation between two variables of *x* and *y* is as follows (Thirumalai, Chandhini and Vaishnavi, 2017):

$$\rho = 1 - \frac{6 * \sum d_i^2}{n\ (n^2\ - 1)} \quad (6)$$

The symbol ρ is denoted as rank relations among the variables, *n* is the number of patients in each group of gender. The d_i is represented as rank difference for variables *x* and *y*. An available SciPy package in python, `scipy.stats.spearmanr` to calculate the spearman's rank correlation is employed (Virtanen *et al.*, 2020). It will give two outputs i) correlation value

between two variables and ii) the p-value to test hypothesis. The correlation coefficient value is interpreted using the rule of thumb in quantifying the strength of association as reported in (Mukaka, 2012).

3.5 Summary

Two case studies were developed using datatset (i) MIMIC-III dataset and (ii) Bradford dataset. The case studies will performed the frailty progression and confirmatory analysis based on the methodology discussed in this chapter. THe methodology used in this work were developed based on three existing methodology, (i) L* life cycle, (ii) PM2 and (iii) MEDCP. It comprises of five stages: (1) planning, (2) Extraction, (3) Data Transformation and Loading, (4) Mining & Analysis and (5) Evaluation. The next three analysis chapter will discuss the frailty progression and confirmatory analysis using MIMIC-III and Bradford dataset. In Chapter 4 describes the preliminary work with frail elderly and frailty trajectories using MIMIC-III dataset, Chapter 6 discuss the frailty trajectories and progression using Bradford dataset and finally in Chapter 7 presents the confirmatory analysis.

Chapter 4

Preliminary Study: MIMIC-III Dataset

This chapter will discuss our preliminary work with the MIMIC-III dataset as our first case study. We begin by explaining how the event log from the MIMIC-III database is acquired and undergone processing and data transformation for analysis. Later, experiments are conducted within our study cohorts. The experiments will investigate the hospital flow, frailty trajectories and progression within the frailty categories. Each of the experiments will be explored in more detail in this chapter.

4.1 MIMIC-III Dataset

4.1.1 Data Provenance and Study Setting

The first EHR data called the Medical Information Mart for intensive Care, release version 3, MIMIC-III (Johnson *et al.*, 2016). The MIMIC-III is a free available data from the Beth Deaconess Medical Centre (BIDMC) in Boston, USA. Boston is the most populated city in Massachusetts and 21st most populated city in the United States with estimated of 692,600 population in July 2019 (U.S. Census Bureau, 2019).

The BIDMC was a union between Beth Israel Hospital and New England Deaconess Hospital about decades ago comprises of more than 4,000 physicians and 35,000 employees. It is a world-leading teaching hospital of Harvard Medical School consists of 673 licensed beds from medical/surgical beds (493), critical care (77) and OB/GYN (62) (BIDMC, 2020). The MIMIC-III dataset includes patient admitted to critical care unit between 2001 and 2008.

4.1.2 General Format of the Tertiary Care

The tertiary care also known as critical care unit which is part of the three levels of medical care. It defined as “care of highly technical and specialised nature,

provided in a medical centre, usually one affiliated with a university, for patients with unusually severe, complex, or uncommon health problems" by the US National Library of Medicine's Medicinal Subject Heading thesaurus in 2013. When it used to define the care given by professional healthcare to patients it usually involves the multimorbidity or the combination of person's viewpoint such as family situations and personality (Flegel, 2015).

4.1.2.1 International Classification of Disease

The International Classification of Disease Ninth revision (ICD-9) codes is the official clinical coding and classification system in the United States. The code had been active in use from 1979 until 1998 before being replaced by the tenth revision of code in 1999 until now (U.S Department of Health & Human Services, 2015). The MIMIC-III data employs ICD-9 to allocate codes to diagnoses and procedures related to hospital utilisation.

The ICD-9 is a numerical code ranges from three to five digits comprises of list of diagnoses code in a tabular format, diseases entries ordered in alphabetical index and a classification system for diagnostic, surgical and procedures. There are approximately 14,500 diagnosis codes and about 3,800 of procedures codes. The five digits structural format of the codes consists of category for the first three digit (where the first digit is either numeric or alphabet E or V), a period as a separator "." and the etiology, manifestation and anatomic sites for the rest of the digits (Cuadrado, 2019).

4.1.3 Data Characteristics and Selection

The access to the MIMIC-III was granted after completing the training course online. We choose MIMIC-III as our first dataset as it openly available and supporting reproducibility of the research using the dataset. The total patient of 46,520 in the data covers the laboratory measurements, charted observation, clinicians' notes, echocardiography and electrocardiogram reports during the hospital stay (Johnson *et al.*, 2016).

The MIMIC-III database created in an open-source database relational management system PostgreSQL. It includes downloading all 26 csv files and import them into the PostgreSQL database using the script available. Figure 4.1 illustrates the ERD for the MIMIC-III dataset comprises of all the tables. The ERD was built using a web-based flowcharts and diagrams creation platform, Lucidchart.

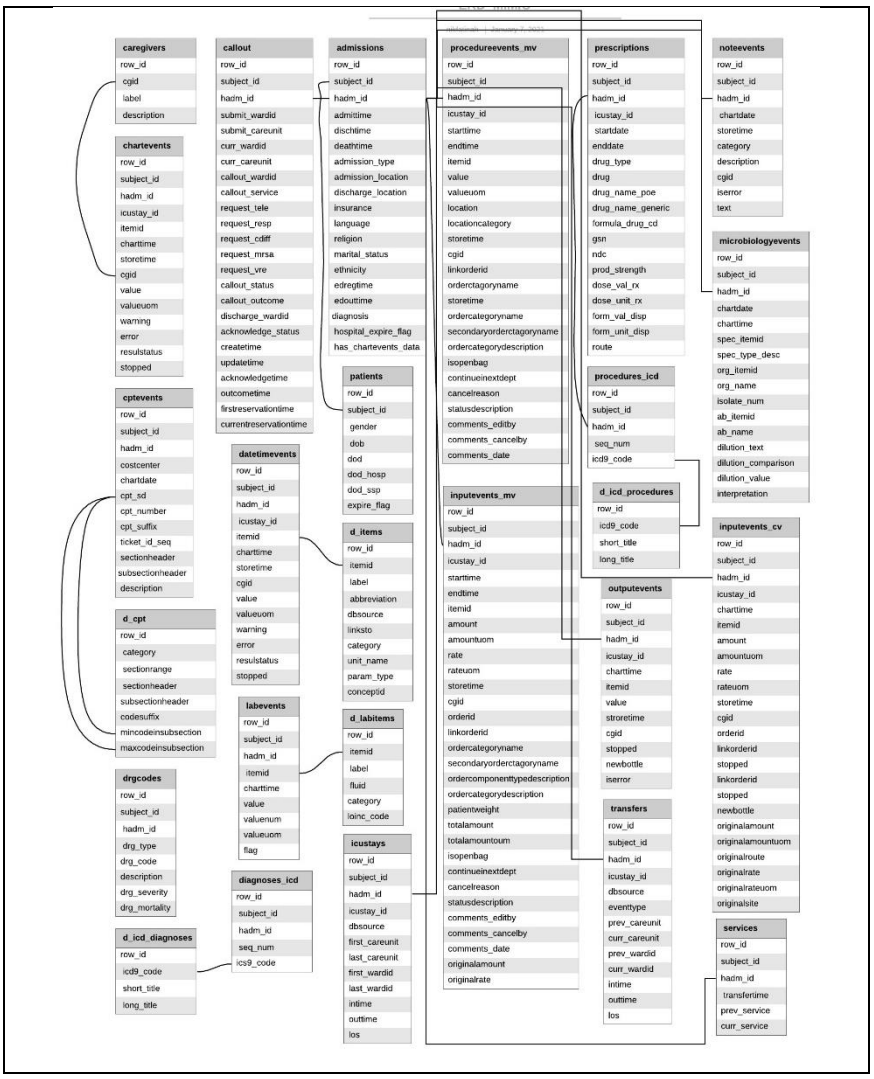


Figure 4. 1 The Entity-Relationship Diagram (ERD) of the MIMIC-III Dataset.

The patients with aged 65 years and above during the admission date were selected using the *patient, admission* tables to identify the age and the date of admission to hospital. However, patients aged over 89 years were kept hidden by shifting their age to 300 years in the MIMIC-III (Johnson *et al.*, 2016) to follow the Health Insurance Portability and Accountability Act, HIPAA regulations.

4.1.4 Event Log Extraction

The standard event log for the modern business process usually is automatically generated from any information system. However, there is some information system, including healthcare required a systematic approach in extracting event logs due to the implicit process formation. An overview of the extraction method for the event log is illustrated in Figure 4.2 as a guide for event log extraction.

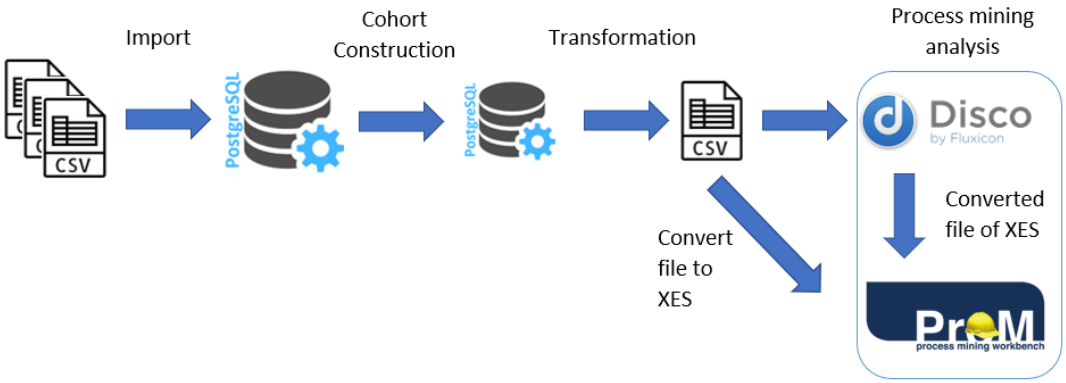


Figure 4. 2 The overview of the event log extraction of the MIMIC-III Dataset.

The event log creation for MIMIC-III database was done in several steps before analysis. It first starts with cohort construction within the PostgreSQL environment to cover two groups of elderly patients; old group with age 65 until 88 and very old group with aged 89 and above. The cohort data will undergo a transformation process of identifying the frailty category in each group in the Python Jupyter environment in the format of csv file. Each elderly group will have four frailty categories; clinically fit, mild, moderate and severe for process mining analysis in process mining tools. The identification of the frailty category is explained in the data transformation step.

4.1.5 Data Quality Inspection

The fundamental data quality for clinical data research established five aspects of data quality (Weiskopf and Weng, 2013). Later the data quality assessments were enhanced with three aspects and additional two context in (Kahn *et al.*, 2016) with research that uses secondary healthcare data that highlight the distinct context of quality assessment. It based on the evaluation of data to the internal characteristics and the external sources. The data quality inspection was followed accordingly to fulfil the requirement for process mining project analysis.

Three data qualities demonstrated in (Kahn *et al.*, 2016) is applied to assess the quality of the MIMIC-II database. The first category of data quality inspection will determine the (1) completeness of the value of data at present, (2) conformance of the value of data to the standardised format and (3) plausibility to identify whether the data can be trusted. Different strategies will be used to assess each context which divided into verification and validation. The validation step will involve using a cross check approach to identify whether our finding agrees with what reported in the MIMIC-III schema analysis data descriptor. The schema

analysis data descriptor is available in https://mit-lcp.github.io/mimic-schema-spy/columns.byTable.html.

4.1.5.1 Data Quality Category I: Completeness

The first category of data quality will focus on the completeness of the data. The verification step was done following to the minimum requirement for process mining analysis which need to have at least user ID, activity and timestamp. The primary user id is *subject_id* present in 16 tables. The second user id is *hadm_id* which record the unique number of admissions into hospital also exist in 16 tables. Another additional user id was *icustay_id* represents the unique number of ICU stay which available in seven tables. The number of user id indicate the granularity of the event for process mining during analysis as a patient will have more than one admission id and ICU id.

The activity obtained without reference from other tables present in all 16 tables. Whereby nine tables require reference from the *d_items* and *d_labitems* table to acquire the activity attribute. The level of specificity for timestamp stored with either date only or with the combination of date and time. The *charttime* is when certain observation is made, and it is validated with the *storetime*. It can be divided into two types; single and couple depending on the level of analysis required. The single type is when only single timestamp recorded for each activity entries and couple have both start and end timestamp. The couple type of timestamp indicates that the duration of event analysis can be performed accurately for a single activity given. Table 4.1 presents the list of tables in each attribute.

Table 4. 1 Summary of the completeness aspect for each required attribute for process mining

Attributes Required	Name of Table
User ID (subject_id, hadm_id)	*admissions, patients, callout, icustays, transfers, services, cptevents, microbiologyevents, charteveents, caregivers, prescriptions, noteevents, inputevents_cv, inputevents_mv, labevents, procedureevents_mv*
User ID (icustay_id)	*admissions, callout, cptevents, icustays, noteevents, prescriptions, microbiologyevents, services and transfers*
Activity (without reference)	*admissions, cptevents, callout, prescriptions, icustays, noteevents, microbiologyevents, transfers, services*
Activity (with reference)	*chartevents, inputevents_cv, outputevents, inputevents_mv, procedureevents_mv, datetimevents*
Timestamp (Single)	*admissions, callout, procedureevents_mv, prescriptions, noteevents, microbiologyevents, chartevents, cptevents, labevents, inputevents_mv, inputevents_cv, outputevents, transfers, services, icustays, datetimevents*
Timestamp (Couple)	*icustays, transfers, procedureevents_mv, inputevents_mv*

4.1.5.2 Data Quality Category II: Conformance

The conformance category of data quality concentrates on the compliance of the data representation according to the set standard and format. The first subcategories assess the value of gender attributes. Our finding conform that all gender value stored in the correct format of character value of ‘M’ or ‘F’. The date shifting value was assessed where we found that not all date is shifted between year 2100 to 2200. The inconsistency with the data descriptor on the date shifting issue may be due to the historical events recorded such as for the scheduled treatments.

The value conformance is the next subcategory where the relatedness and coherence between tables is assessed for the accuracy measure is called relation conformance. Both *d_icd_diagnosis* and *d_icd_procedures* had less than 1% missing code (144 out of 14,711) and (16 out of 258,082). The redundancy for date of admission to hospital was identified in tables *admissions* and *datetimeveents*. However, the *admission* table was selected for the date of admission as 92% (22,658 out of 24,549) had earlier admission date compared to the *datetimeevents* table. The computational conformance assessed the eFI scores calculated using previously explained in (Clegg *et al.*, 2016). The detail description of the eFI score is presented in the section 1.1.6.2 which is part of our data preparation step.

4.1.5.3 Data Quality Category III: Plausibility

The last data quality category is focuses on defining the credibility of the data. The two subcategories of plausibility investigate the uniqueness and the atemporal aspect of quality. The uniqueness seeks to determine the consistency used in the icd9 codes selected for each frailty deficits where no code should be in two different deficits. The atemporal determine the distribution and density of the observed group whether it agrees with the external sources. The detail of the uniqueness and atemporal data quality will be explained in detail in section 1.1.6.1 later. The last quality aspect concentrates on the temporal properties of time-varying variables where value should match with the establish expectation. The admission and discharge date in the `admissions` table is compared. We found that the discharge date always be on the same date as admission or at least a day after the admission date which conform to the established medical data (Nanayakkara *et al.*, 2014).

4.1.6 Data Transformation

In this section, we will discuss data transformation in detail. It is a preparation step to create event logs. Data preparation was done following the logging guidelines which emphasize the combination of both the transformation process and accessibility to the raw event data (Van der Aalst, 2016). However, the creation of event log from raw event data requires selection of relevant event to the process, correlated events to form process case, explicit ordered event using the timestamp information and events attributes need to be computed from the raw data. The approach of adopting the Guideline Logging (GL) will be explained in detail along in this section.

Although it is not formally mention in this study, the existence of the healthcare activities extracted from the EHR consider representing the transactions information (GL7). It is importance to highlight that MIMIC-III database performed a structured data cleansing approach in ensuring the patient privacy (GL12) according to the HIPAA. The experimental sheet was well documented to ensure the syntactical correctness ad consistency of the event log used throughout the work which is part of the logging guideline (GL8).

4.1.6.1 Dealing with Temporality Issue

There are events recorded with different date time format such event with only date and other event includes with timestamps details of hours, minutes and seconds. An example of such event are events stored in the `prescriptions`

table in MIMIC-III database. The different granularity of timestamp will create a strong influence to the process model produced using any process mining tools (Kurniati *et al.*, 2019). It will produce a misleading process model as the real order of the events is inconsistent. However, to address this matter within our work, we use the higher granularity of timestamp stored in the MIMIC-III database, date with the timestamp details of hour, minutes and seconds. We also ought to update the timestamp detail to the events which timestamp detail is absence to 00:00:00. The improvement steps taken following the logging guideline to ensure the quality of the event log is at optimum level. The logging guideline involved are attributes value needs to be as precise as possible (GL4) and implicitly ensuring the order of the events such as using attributes representing the timestamp as the events order (GL6).

4.1.6.2 Electronic Frailty Index Score

The electronic frailty index (eFI) score is an instrument to help identify the elderly patient who has high susceptibility to adverse health outcomes in primary care (Lansbury *et al.*, 2017). It based on the cumulative deficit model which uses the accumulation of diseases, symptoms, disability, and laboratory test abnormality from the EHR as the measurement of frailty called deficit developed by (Clegg *et al.*, 2016).

The frailty index score is representing the total count of deficit presents out of the total deficits recorded within the study duration. Next, the frailty score is used to classify the severity of frailty within the elderly patient. The classification grouped the frailty severity into four categories: clinically fit, mild, moderate, and severe. We follow the estimates prevalence of 50%, 35%, 12% and 3% respectively for each categories provided in (Clegg *et al.*, 2016).

The ICD9 codes associated with frailty following the approach described in previous study was manually selected (Clegg *et al.*, 2016). A total of 905 codes was grouped into each individual deficit. Due to the nature of the healthcare setting within the tertiary care, no deficits such as activity limitation, housebound, requirement for care and social vulnerability were identified in the elderly cohort. Additionally, the polypharmacy deficit also was omitted after evaluating the `prescriptions` table, that the duration of medications prescribed only within the hospital stay. No additional data provided to determine the patient was in condition of taking medication after the stay.

4.1.7 Data Profiling

The data profiling describes the data consists in MIMIC-III related to the elderly patient in the database population for the study. The patients were included with a main criterion of at least one admission into the any critical care. Table 1.2 shows the distribution of the disease and injuries codes among different gender in elderly population. It shows MIMIC-III dataset represents a variety of disease diagnosis. However, since the study population is elderly over the age of 65 years, no chapter code from complication of pregnancy & childbirth and conditions in perinatal period id recorded. It appears both genders had the similar top three diagnosis code recorded which are highlighted in the Table 4.2.

Table 4. 2 Distribution of the ICD9 codes groups within the elderly patient. The number represents the total number of diagnosis in each chapter code, where a patient might have more than one diagnosis. The percentage is calculated over each gender population.

Chapter Code	Description	Male (n=10,304)	Female (n=9,460)
001-139	Infectious disease	3,913(31%)	3,835(28%)
140-239	Neoplasms	2,502(17%)	3,221(22%)
240-279	Endocrine, metabolic disorder	**14,049(79%)**	**14,471(77%)**
280-289	Disease of blood & blood forming disease	5,428(44%)	5,663(41%)
290-319	Mental disorder	3,924(31%)	3,511(26%)
320-389	Disease of nervous system	4,467(32%)	4,747(31%)
390-459	Disease of circulatory system	**31,866(95%)**	**36,926(95%)**
460-519	Disease of respiratory disease	9,983(56%)	10,689(54%)
520-579	Disease of digestive system	7,547(45%)	7,970(43%)
580-629	Disease of genitourinary system	7,537(50%)	9,505(55%)
630-679	Complication of pregnancy & childbirth	0(0%)	0(0%)
680-709	Disease of skin tissue	1,417(13%)	1,598(13%)
710-739	Disease of musculoskeletal system	4,036(31%)	2,350(18%)
740-759	Congenital anomalies	212(2%)	285(3%)
760-779	Conditions in perinatal period	0(0%)	0(0%)
780-799	Symptoms, signs conditions	6,208(43%)	6,885(43%)
800-999	Injury and poisoning	7,258(48%)	8,290(50%)
E800-E999	Classification influencing health status	4,361(35%)	4,616(34%)
V01-V89	Classification of external causes of injury and poisoning	**10,705(58%)**	**12,969(62%)**

4.2 Experiment 1: The hospital's flow within different frailty categories

This section describes the first experiment conducted for our preliminary study. The main aim for this experiment is to investigate the suitability of applying process mining for analysis. The general methodology discussed in Chapter 3 is

implemented while carrying out the experiment. The complete documentation of this experiment is outlined in Appendix A.4.

4.2.1 Stage I: Planning

The initial stage started with planning the experiment and developing research questions for the analysis. The general research question in this experiment is '*Can process mining techniques be applied into frail elderly MIMIC-III dataset to analyse the hospital workflow?*'. The main reference for the MIMIC-III database for data descriptions and data schema is our main reference in understanding the database (Johnson *et al.*, 2016). We constructed our detail research questions following the frequently asked questions by the healthcare professional (Mans *et al.*, 2013) as follows:

i. What is the most followed path and the exceptional path of the frail elderly patient?
ii. Is there difference in path followed by different frailty categories?
iii. Is there any bottleneck in the path of frail elderly patient?

The detailed research questions above addressed the main RQs of the research, (RQ1) "Is it possible to analyse frail elderly pathway using data and process mining analysis?", (RQ2) "What is the best illustration to portray frailty pathway?" and (RQ4) "How can the dataset in relation to eFi scores be extracted from the EHR system?". The general question in this experiment explored the (RQ1) and (RQ2) to find the most followed and exceptional path, the difference in path followed by different frailty categories and any bottleneck. (RQ2) is addressed to determine the best visualisation to present the path. These can be achieved by exploring the (RQ4) by extracting the frail elderly dataset based on the frailty categories in the MIMIC-III dataset.

4.2.2 Stage II: Extraction

The elderly patient is selected based on the age of 65 years and above from their first admission into hospital. The selection is based on the two tables: `patients` and `admissions`. Figure 4.3 shows the patient distribution of two-years age range. The extraction obtained a total of 19,764 elderly patient within the MIMIC-III database. There were 100,41 (50.8%) of male patient and 9,723 (49.2%) of female patient, with highest percentage came from the patient age group of 89

and above 300, 1,986 (10.26%). The number agrees with the main reference as our MIMIC-III data descriptor for checking purpose.

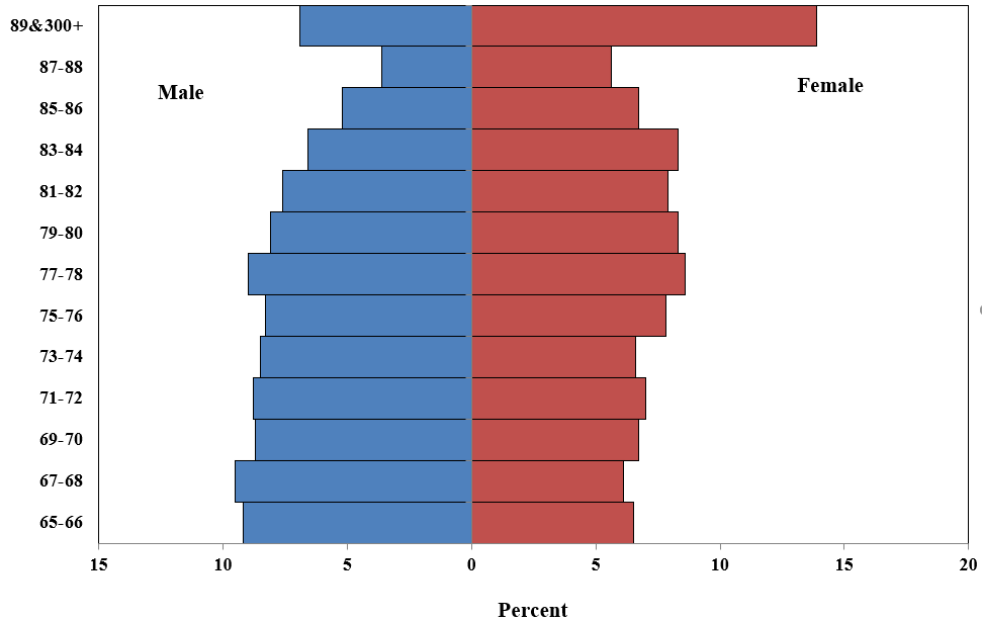


Figure 4. 3 The age distribution of the elderly patient during the first admission. The x-axis shows the gender and age range group percentage of patient within the elderly population and the y-axis shows the two-years age range. The MIMIC-III has changed the patient with age more than 89 years old to over 300 to comply with HIPPA rules to keep the patient confidentiality. The gender is coded into blue to indicate male (on the left side) and red to indicate female (on the right side).

The records related to elderly patient extracted in previous stage was obtained from the 17 tables in the MIMIC-III database. The list of tables covers all records of patient is present in Table 4.3 along with number of unique patients (n), activity's column name, unique activities and total rows for each table. The top three biggest tables are *chartevents*, *callout* and *inputevents_mv*. The total events for the elderly patient in the MIMIC-III dataset is too large (~72 million) which will produce a very complex process model. Thus, further process transformation is crucial for the process mining analysis to yield an understandable model.

Table 4. 3 Summary of the records extracted for the elderly patients

Name of Table	n	Activities	Rows
admissions	19,764	5	25,511
icustays	19,764	2	26,972
services	19,761	18	32,934
transfers	19,764	5	114,917
diagnoses_icd	19,764	4,583	326,043
prescriptions	18,231	3,003	1,720,719
chartevents	19,752	2,746	**45,876,379**
labevents	19,719	678	**10,737,715**
callout	11,560	18	17,199
cptevents	17,461	1,348	271,880
datetimeevents	14,140	153	2,159,909
notevents	19,261	15	726,754
outputevents	19,559	616	2,168,056
procedurevents_mv	9,059	12	126,322
inputvenets_mv	9,050	3	1,624,303
inputvenets_mv	11,378	260	**5,394,778**
microbiologyevents	16,977	77	161,292
Total	-	13,542	71,511,683

4.2.3 Stage III: Data Transformation and Loading

The third stage of the experiment is **data transformation** and loading. It was performed on the raw or extracted data event with total of 71,511,683 rows. The transformation is part of data processing with aim to group and select event type for creating event log which incorporates the transactional information. The definition of transactional information is events contain within a business unit such as ‘start’, ‘end’, ‘suspend’, ‘resume’, ‘complete’ and etc (van der Aalst, 2015). In the healthcare context transactional information is sequence of records associated with patient flow within the hospital. The transformation following the process mining project methodology (Van Eck *et al.*, 2015) consists of three steps, filtering log, creating view and enriching logs. The detail of data transformation is illustrated in Table 4.4 with percentage calculated from the total rows of the extracted data at each data transformation steps.

Table 4. 4 Data transformation steps; filtering log, creating views, and enriching log

Step	# rows	Percentage (%)
(0) Extracted Log	71,511,683	100.00
(1) Creating Views	2,483,439	3.47
(2) Filtering Log	188,804	0.26
(3) Enriching Log:	**Frailty Category**	
Mild	63,049	< 0.0001
Moderate	34,433	< 0.0001
Severe	7,077	< 0.0000001

The first data transformation is (1) creating view where an event log is created by specifying and selecting the view of events data which heavily depends on goal

of the experiment. The event classes within the event data is determined to distinguish different level of activities. The administrative event, clinical events procedure events, prescriptions events are the main event classes. As the goal of this experiment to investigate the hospital workflow of patient, only administrative event with a total row of 2,483,439 (3.47%) is selected. The next step is (2) filtering logs where it deals with the duplicate events by keeping only one of them to reduce the complexity of the analysis. The number of rows left is 188,804 (0.26%) from the extracted data. The next step is (3) enriching logs where a frailty category is added as an additional event. Deriving the frailty category is following the steps describes in section 1.1.6.2. The last step of data transformation is also the next iteration of log filtering is done where the log is sliced based on its frailty category.

The final step is **loading** the processed event log into the process mining tools; Disco and ProM for analysis to discover process models.

4.2.4 Stage IV: Mining and Analysis

Mining and analysis were the next stage of this study. The processed event log is used to conduct two main types of process mining to answer research questions. In addition to that, process analytics approach is implemented to compare frail pathways of different frailty categories. The fuzzy miner algorithm is used on the event log as the preliminary algorithm to answer general research question which is "*Can process mining techniques be applied into frail elderly MIMIC-III dataset to analyse the hospital workflow?*". The explanation of how to interpret the model has been discussed in Chapter 2 in Section 3.2.4.1.2. The generated process model is analysed to observe whether it really portrays the reality of hospital flow within frail elderly. Processed event logs are initially mined using the Disco application to generate the XES files to be further analysed using ProM application.

Different type of visualisations will be used to answer the first research question (i) "*What is the most followed path and the exceptional path of the frail elderly patient?*".

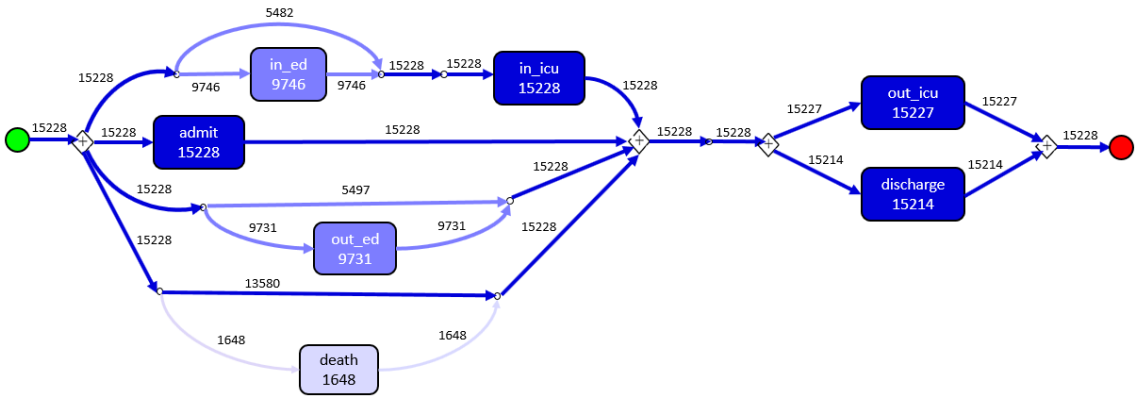


Figure 4. 4 Process tree produced from the IvM plug-in in ProM. The frail pathway within the hospital admission.

The Process Tree using the Inductive Miner (IvM) plug-in in ProM is produced using the default setting and illustrated in Figure 4.4. The model is a representative of 15,228 admissions with 10,512 elderly patients within the hospital admission and into Intensive Care Unit (ICU). The model presented to clinicians based in UK. It is interesting to observe that the admission pathway could end either with discharged from hospital or out of the ICU.

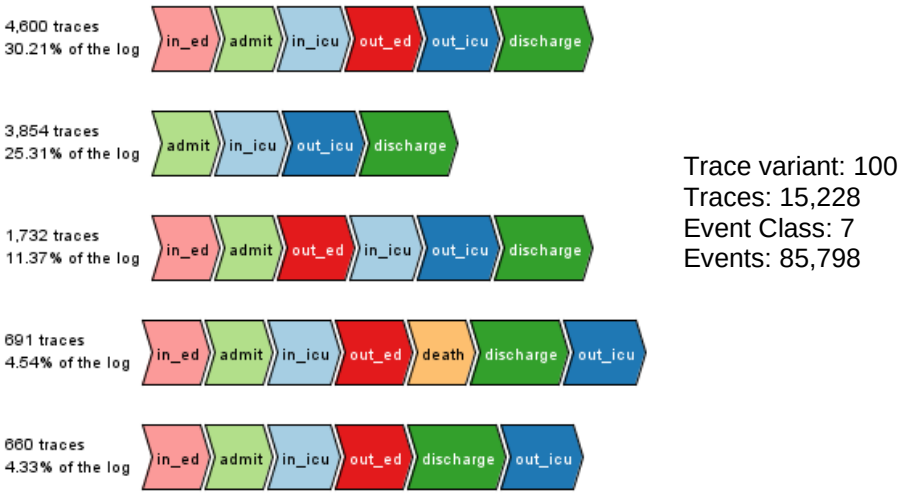


Figure 4. 5 Top five common traces generated in ProM. It covers 5% of all trace variants

Top five trace variants are observed to answer research question (i) in Figure 4.5 as another visualisation. The trace variants are from a total of 15,228 traces. They covered 11,537 (75.8%) of all traces imitated the hospital admission flow as part of their administrative process with numbers of characteristics. The flow of admission could start from registering to the Emergency Department (ED) or direct to hospital admission until being discharged. A compelling pattern was observed that some patients could be out of the ICU (`out_icu`) even after been

discharged (`discharge`) from the hospital (*the fourth and fifth trace variant from the Figure 1.4*) which could be considered as data quality and accuracy issue.

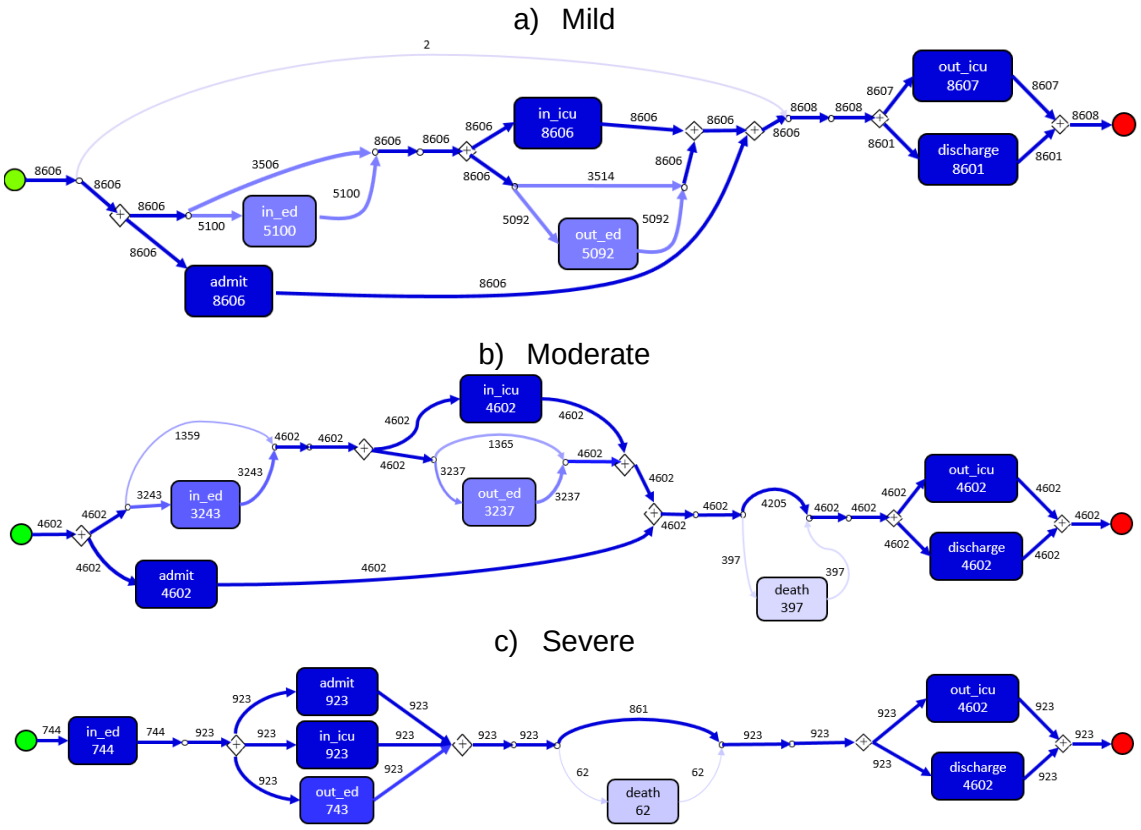


Figure 4. 6 Pathway of each frailty categories in process trees in ProM.

The second research question (ii) "*Is there difference in path followed by different frailty categories?*" is answered next. Process models of frail patients from three categories; mild, moderate, and severe are generated following the previous method. The model represents the significant pathway of each frailty categories in Figure 4.6. The models undergo visual evaluation to identify any differences and similarities between them. Some finding on similarity between models of all categories are: 1) despite of the activity ordering, six activities; admit, discharge, `in_icu`, `out_icu`, `in_ed`, `out_ed` always happened, 2) four activities always happened in sequential order; `admit` happened before `discharge` and `in_icu` happened before `out_icu`, 3) the process could end either by `discharge` or `out_icu` activities. Furthermore, the differences between categories are observed such that 1) death is only possible in moderate and severe before `discharge` and `out_icu`, 2) in severe the process starts by in_ed activity, 3) unlike in mild and moderate where either activities `admit` or `in_icu` as the start of the process.

The final question of the experiment is determining the bottleneck of admission process of frail elderly patient. The bottleneck of the process is where an activity having the longest time to complete before executing the next activity. The path and sojourn time is presented as the visualisation mode as shown in Figure 4.7. It is generated from the 'show' features available in the plugin.

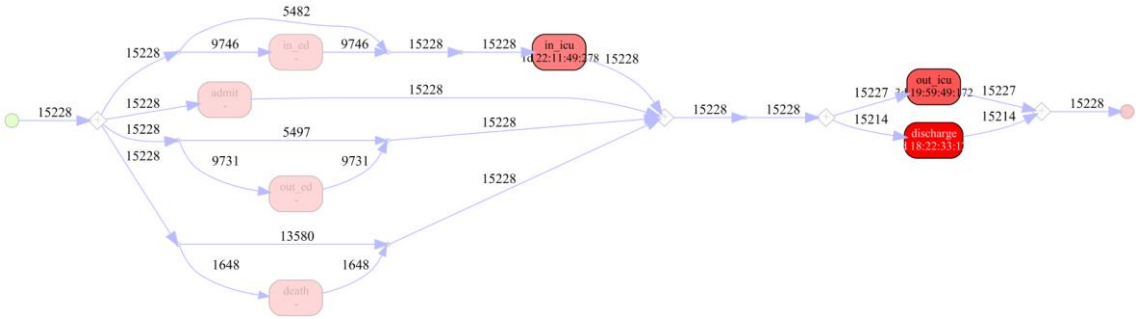


Figure 4. 7 Path and sojourn time process tree model generated by IvM plugin in ProM

The analysed process model is from process tree of the first question in Figure 4.7. The activity is illustrated in red box with higher intensity colour indicates higher average sojourn time. The sojourn time is the time between completion of previous activity and completion of current activity in sequential order. The analysis revealed that activity discharge on average took 7 days, 18 hours, and 22 minutes between the completion of the in_icu, admit, out_icu and death and to end the process. It is the longest sojourn time in the model, while the second highest is out_icu activity. It took an average 3 days, 19 hours, and 59 minutes of sojourn time before ending the process.

4.2.5 Stage V: Evaluation

The suitability of applying process mining into MIMIC-III is the first aspect of evaluation in this experiment. The possibility of creating event log for process mining analysis from MIMIC-III was achieved using approach presented in Section 1.2.3. The second aspect of evaluation is discussion of the finding with UK healthcare professional. In the viewpoint of clinician, the visualization presented as a workflow of network graph was useful as it able to display the whole process of elderly admission within the tertiary care either through emergency department or non-emergency entry. Besides, the process presented were easy to understand by any healthcare professionals and beneficial in analyzing the variation of the process. However, the main concern in this part is limited knowledge UK healthcare professional has regarding the US healthcare system. The model checking based on common logic following the common

healthcare workflow is the only way to perform which can pose concern on possible flaws with US healthcare system.

4.3 Experiment 2: Frailty Trajectories within Different Categories

The purpose of the second experiment is to examine the suitability of combining the process mining and eFi with routinely collected record to analyse the variability of frailty trajectories. Frail patient is selected as the cohort for this experiment and their respective deficits associated with frailty. Each steps of the general methodology implementation are discussed in this section and outlilned in experiment documentation in Appendic A.5.

4.3.1 Stage I: Planning

The second stage was done to plan the experiment. This is an initial experiment in assessing the frailty progression. The primary research question is that “Can process mining be used to explore the variation in frailty deficit pathway within frailty categories?”. The analysis focused on the progress of diagnosis related to frailty. This associates with (RQ1), (RQ2), (RQ3) and (RQ4) of the research. It aims in determining the possibility of applying process mining techniques in studying frailty progression with the best illustration and using a sophisticated approach in extracting the dataset for the experiment as will be discuss in the next section 4.3.2. The development of frailty is considered as a sequence of diagnosis identified since the start until last hospital admission. The experiment aimed to explore the variation using the model or network as the representational trajectories within frail elderly.

4.3.2 Stage II: Extraction

The extraction of cohort is done in two steps. The first step is selecting cohort based on first admission to hospital with minimum age of at least 65 years old. The selection was referred to tables; *patients*, *admissions*, *diagnoses_icd* and *d_icd_diagnoses*. The age of patient was retrieved based on the date of birth in patient table and their first admission in *admission* table. The other two elements for process mining analysis was derived from the tables: *admissions* for the timestamps, *diagnoses_icd* for the activity (as frailty deficit) which referred to the *d_icd_diagnoses* as code reference. The timestamp for deficits were recorded based on their date and time

of each admissions to hospital. The second step of data extraction was to only include patient with number of admissions more than once.

4.3.3 Stage III: Data Processing and Transformation

The aim of this stage is to create event log from the extracted event data. It can be achieved by transforming the data to be ideal for analysis following the process mining project methodology step (Van Eck *et al.*, 2015). The transformation steps are (1) log enriching, (2) log filtering (3) creating view and (4) aggregating the events. The first step is to enrich the log by computing additional events.

The 1) log enriching step was done to determine the patient frailty index score and their frailty category during each admission as explained in the section 1.1.6.2. All ICD codes (`diagnoses_icd` and `d_icd_diagnoses`) related to the elderly patients were identified. The codes were selected to be grouped into 36 identified deficit associated with frailty according to previous study (Clegg *et al.*, 2016). The codes grouped into each frailty deficit is presented in Table 4.5 with a total of 905 codes identified within the elderly patient. The highest codes identified is from the fragility fracture deficit, following with the heart valve disease. It is worth to note that no codes recorded from the deficits; activity limitation, housebound, requirement for care and social vulnerability is unavailable within the elderly patient of tertiary care setting.

Table 4. 5 The distribution of the number of codes associated in each frailty deficit in MIMIC-III dataset

DEFICIT	# CODES	DEFICIT	# CODES
Activity Limitation	0	Ischaemic heart disease	43
Anaemia Haematinic Deficiency	47	Memory/Cognitive Problems	11
Arthritis	43	Mobility/Transport Problems	2
Atrial Fibrillation	2	Osteoporosis	2
Cerebrovascular Disease	32	Parkinsonism and Tremor	2
Chronic Kidney Disease	13	Peptic Ulcer	34
Diabetes	46	Peripheral Vascular Disease	11
Dizziness	5	Respiratory Disease	34
Dyspnoea	1	Requirement for Care	0
Fall	31	Sleep Disturbance	11
Foot Problems	10	Skin Ulcer	24
Fragility Fractures	281	Social Vulnerability	0
Heart Failure	24	Thyroid Disease	25
Hearing Impairment	9	Urinary Incontinence	10
Housebound	0	Urinary System Disease	7
Heart Valve Disease	76	Visual Impairment	56
Hypertension	4	Weight Loss & Anorexia	2
Hypotension/Syncope	7		
Sub-Total	**631**	**Sub-Total**	**274**

The second step is filtering the log was performed to reduce the complexity of the process model produced. The filtering technique used was slicing the event log based on their final frailty index score. The 2) log filtering will create four event logs according to patient total cumulative frailty deficit to form frailty categories of clinically fit, mild, moderate, and severe. The event log from the clinically fit category is excluded in this experiment as we only interested in finding out the process flow of elderly patient. Next, determining the type of events and case notion to form process instance was done in the next step of 3) creating view. The type of process instances in this type of study would be created from the clinical or administrative events. Finally, 4) aggregating the similar event and select the first event was done. All transformation step was done using the Python Jupyter notebook platform. Table 4.6 shows the detail of the transformation steps taken. It shows that the transformation steps have reduced more than 85% from the second step of filtering event log in all frailty categories.

Table 4. 6 A detail of the transformation steps with the number of events at each step

Transformation Step	Number of events		
1) Log enriching	378,526		
Frailty Category	**Mild**	**Moderate**	**Severe**
2) Log filtering	66,242	105,988	92,739
3) Creating views	27,593	59,793	26,326
4) Aggregating events	8,137	12,403	3,306

The distribution of the number of admissions with total cumulative deficit of final event logs was plotted as in Figure 4.8. It shows most of the cases in all categories had number of admissions between two to three, but the maximum number of admissions varies between categories. This distribution gives good indication about the skewedness of the cohort which reveals the likelihood of higher number of admissions with more advance frailty condition.

Finally, the event logs from the three frailty categories was loaded into the process mining tool for mining. It was done by exporting the file as the csv format from the Jupyter notebook and importing the event log in the type of XES file using the Disco software.

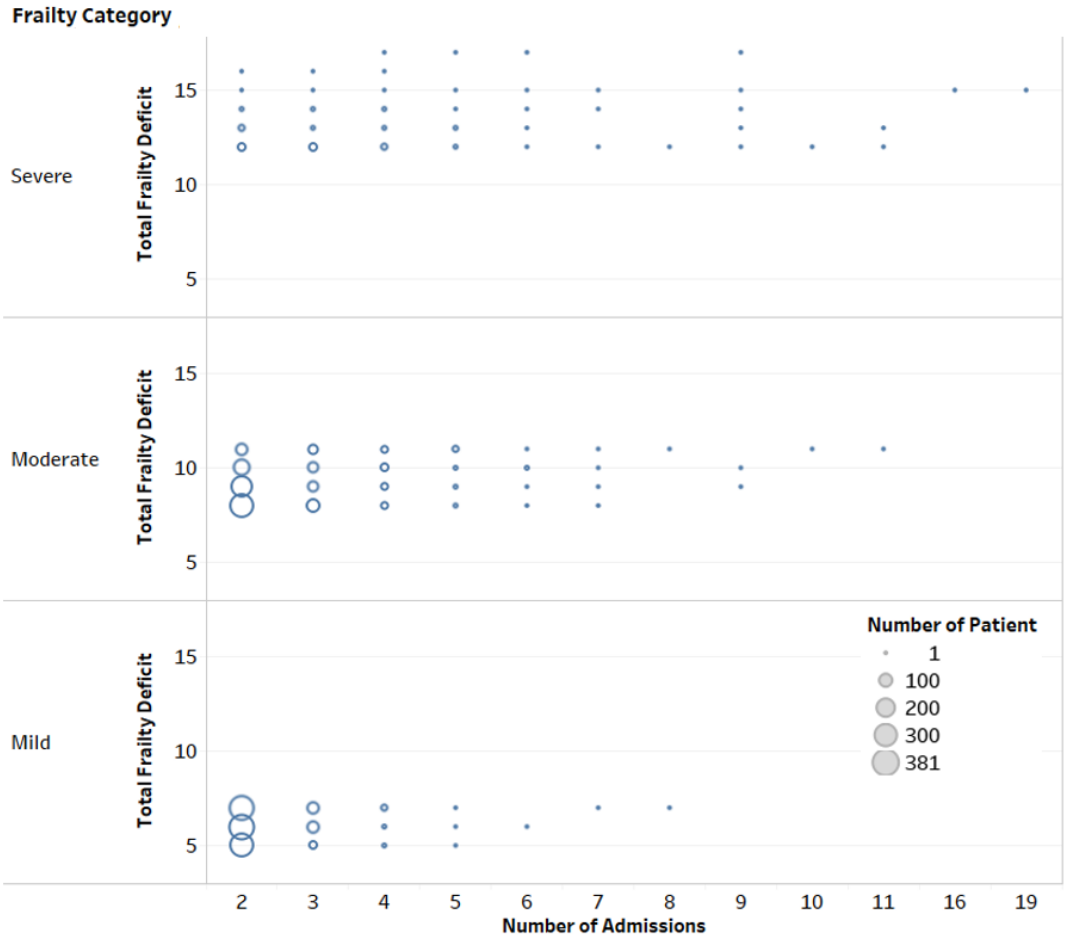


Figure 4. 8 The dotted chart of distribution of number of admissions and total cumulative frailty deficits. The size of the dots represents the number of cases or patients resides in number of admissions and total cumulative deficit within frailty categories. The categories were labelled in the left part of figure.

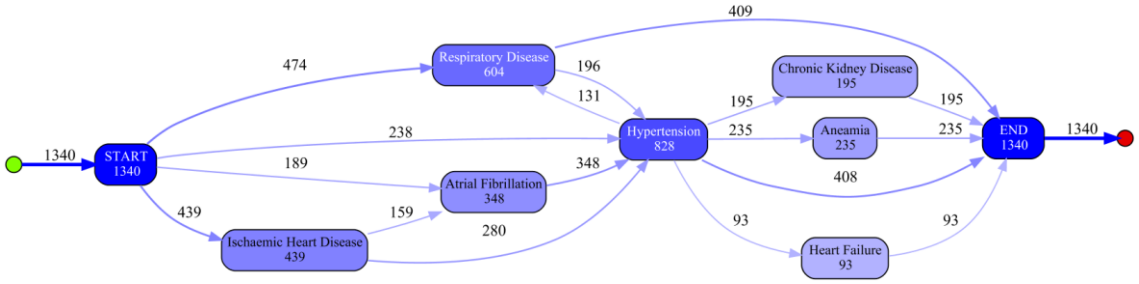
4.3.4 Stage IV: Mining and Analysis

In fourth stage of methodology is mining and analysis where the main analysis involved is process mining through process discovery and conformance checking. The process discovery was done using the Directly Followed visual Miner (DFvM) plugin in ProM.

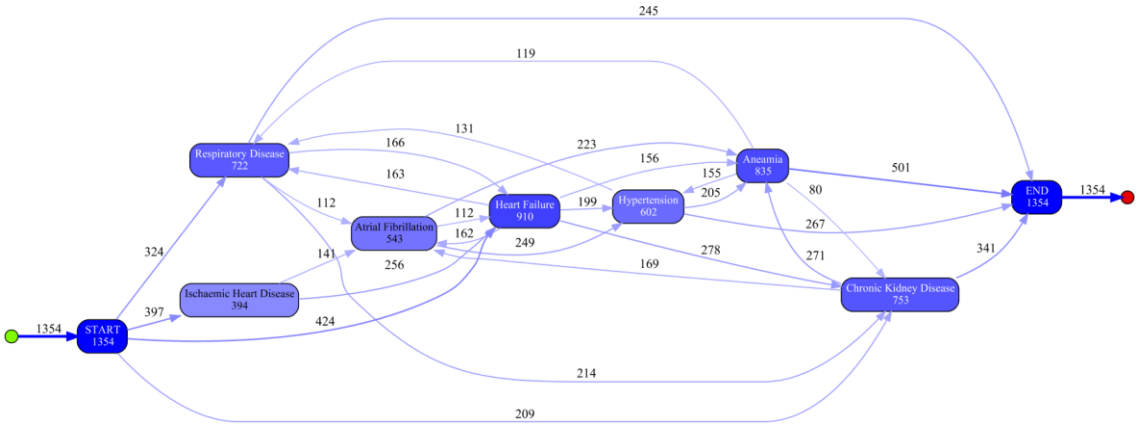
1) Process model as the representational of frailty trajectories

Process models resulted as directly followed graph are illustrated in Figure 4.9. The model describes the directly followed of one event into another event successively. For example, given a list of event A, B, C and D, where the trace of a case is A > B, C > D. The comma indicates the event recorded with similar timestamp. Thus, the directly followed event would be A > B, A > C, B > D and C > D respectively. The plugin was chosen in this experiment as the model provides

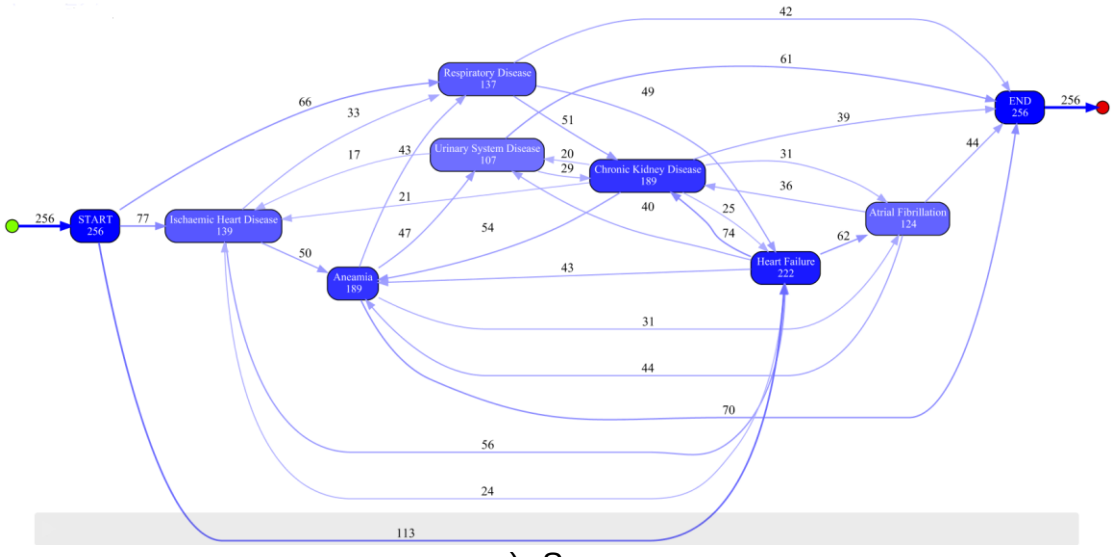
an informative visual of model and adjustable setting mode for the model representation. The box of the models represents the activity and the edges represents the direction or flow of one activity to another. The colour of box represents the frequencies of each activity been executed and the edges is number of cases following the path.



a) Mild



b) Moderate



c) Severe

Figure 4. 9 The directly follows models by plugin DFvM in Prom. The model is set to represent the 25% frequent activities and 1% frequent path.

Figure 4.9 shows the trajectories model across frailty categories; a) mild, b) moderate and c) severe. The manually visual process model comparison approach was done using the created process models using the DFvM plug in ProM. The initial observation with similar setting for all models shows that mild has the least number of edges as compared to the other models which also indicates that high traces share similar variant. The top two start activities in mild are ischaemic heart disease and respiratory disease; in moderate and severe are heart failure and ischaemic heart disease. Hypertension seems to be the centre of the deficit diagnosed after the initial deficit and has the highest total of occurrences, 828 within the traces in mild. While in moderate model, heart failure with total occurrence of 919 received the inflow edges from all initial start deficit and in severe anaemia, urinary system disease, chronic kidney disease and heart failure appear to be centre of deficits to interact between the initial deficits (ischaemic heart disease and respiratory disease) and final deficit (atrial fibrillation) the in model.

2) Process analytics of the trajectories

Further analysis was done to focus on the process metrics logs comparison. It was accomplished through examining the overview statistics of each event logs as illustrated in Table 4.7. The obvious statistics measure among categories can be observed is that severe had the highest event per cases and case duration, both mean and median. The variant percentage score of both moderate and severe had value of 100% indicates that all traces have unique set of sequence activity/deficit. However, in mild only three trace variants shared similar traces with two cases each.

Table 4. 7 Overview statistics of event logs in mild, moderate, and severe. The last column indicates the total variants of each model with the variation percentage score shows in the bracket, where variation percentage is the number of variants divides by total case and times hundred.

Category	# Case	# events	Events per cases	Case duration	Total variants
Mild	1,340	8,137	Min 5, mean 6 and max 7	Median: 4 months and 2 weeks Mean: 1 year 4 months 2 weeks 4 days	1,337 (99.78%) *
Moderate	1,354	12,403	Min 8, mean 9, max 11	Median: 10 months and 3 weeks Mean: 1 year 10 months 1 week and 5 days	1,354 (100.0%) *
Severe	256	3,306	Min 12, mean 13, max 17	Median: 1 year and 10 months Mean: 2 years 8 months 1 week and 5 days	256 (100.0%) *

4.3.5 Stage V: Evaluation

In this experiment, evaluation was done by examining the suitability of determining frailty trajectories with process mining using MIMIC-III dataset. An insight gained from this experiment is that there is good prospect to create frailty trajectories to analyse frailty progression within frail elderly patient. Although the sequence is reliable, this experiment was not possible to present the statistically significant trajectories across frail patient. Another point of insight attained was that the differences between categories could be due to the range of deficits allocated in each category as described in chapter 5 section 5.7.2 which may give concern to the quality of the approach. Although, the concept of representational of process model in DFvM plugin and Disco is similar, the DFvM plugin was chosen over Disco as it produced clear semantic model with similar number of cases represented at the start and end of model (Leemans, Poppe and Wynn, 2019).

An additional evaluation was process conformance checking with process model quality metrics, fitness is measured. Fitness is the degree to which the discovered model can correctly replicate the cases reported in the event log. The fitness score yield for mild is 0.9678, moderate 0.97735 and severe is 0.97268. High fitness observed in all models suggested that all models able to replay all the observed behaviour in the event log.

The clinical feedbacks found that the network of frailty deficits emphasize the frailty trajectories over time is useful in showing the progression of frailty. However, the frailty trajectories presented in this experiment is considered incomplete as the total deficits available in the dataset was twenty-seven out of thirty-six as suggested in the literature. An extension of work in analysing the comprehensive frailty trajectories are suggested to include all thirty-six deficits.

4.4 Summary

This chapter has described the preliminary analysis with MIMIC-III dataset. The case study can be analysed in various perspectives. In process mining perspectives it proved that MIMIC-III can be used for process mining in the healthcare domain based on the process discovery and conformance checking. Besides, the analysis in this chapter provides the healthcare professionals with some insights such as the most followed paths, exceptional paths, the frequency-based questions for healthcare perspective as in experiment 1 in section 4.2. Although, MIMIC-III represents real healthcare data there are elements posed as

limitation when analysing with MIMIC-III data. The main limitation recognised was that all dates have been shifted randomly to next decades years (between 2100 and 2200) consistently. It restricts any real time-based analysis between cases such as network analysis or even waiting time analysis of busy day. Other limitation identified was there was no comprehensive patient data from other clinical setting included such as in primary or secondary care. Despite of the data quality of MIMIC-III dataset found and described in section 4.1.5., it still provides rich set of healthcare dataset for analysis. The works in this chapter offer great opportunity to investigate frailty progression using process mining.

Chapter 5

Acquisition of Event Log: Connected Bradford Dataset

This chapter will discuss the steps in acquiring the event log for the analysis work. It begins with understanding the origin of the data source and the context of the data been generated into the EHR. The raw dataset will undergo several steps of data quality assessment, data preparation and data processing as part of data transformation before selection of study cohort is made.

5.1 Study Setting and Context

The dataset was retrieved from the healthcare system of National Health Service (NHS) England. The NHS is the system of public healthcare in the UK, which includes NHS England, NHS Scotland, NHS Wales, and affiliated Health and Social Care in Northern Ireland. It provides health care support and services divided into cancer, mental health, primary care, integrated care, and urgent and emergency care. Healthcare is provided in two parts, which are primary care that involves General Practitioners (GP) and community services and secondary care, which consists of hospitals and specialists.

The study uses the dataset from the primary care setting. The primary care provides the first point of contact in the healthcare system, which also acts as the 'front door' for the NHS. It includes general practices, dental, pharmacy, and optometry services to the public. It is day-to-day healthcare available in every local area, and the first place the public go when they require treatment and any health advice. The health promotions, disease prevention, health maintenance, counselling, patient education, diagnosis and treatment of acute and chronic illness in a variety of health care settings are several of many things provided by the primary care practices.

Within the UK healthcare setting, individuals seeking advice or treatment for a health concern usually meet with a family physician (known as General Practitioner, or GP) or a nurse (for example, a Nurse Practitioner) at their local general practice. GPs can refer their patient who requires more specialised

treatment (or further test) to the hospital or other community-based services. There is a wealth of information available within primary care records; most secondary care interactions are reported back to general practice, and some illnesses tend to be managed entirely within the primary care setting.

5.2 Data Provenance

The source of dataset for the introductory work was the EHR from Bingley and Saltaire GP Practices, called as the small dataset. The two GP practices are in the northern English county of West Yorkshire in England. Bingley GP practice is a GP surgery and a teaching practice providing care service to almost 12,000 to local community (BMP, 2019). Saltaire medical practice formed a partnership with the Windhill medical practice established in 2018 named as The Saltaire and Windhill Medical Partnership providing access to care to almost 50,000 population (SWMP, 2021). The remaining dataset called full dataset is the combination of 86 GP practices across Bradford including the previous two.

The full dataset was extracted from the hospital information system (HIS), SystmOne of the GP Practices. The extracted dataset sored in the database server of Microsoft SQL Server 2013. It comprises copy of live database of patient record received care services under the GP practices from the year 1877 until the present. In this dataset, it used Read Code version 3 (CTV3) that supports specific clinical encoding of multiple patient phenomena including occupation, social circumstances; ethnicity and religion; clinical signs, symptoms and observations; laboratory tests and result; diagnoses; diagnostic, therapeutic or surgical procedures performed; and a variety of administrative items.

The access to the GP practices across Bradford was provided by the Bradford Institute of Health Research (BIHR). The right to conduct research was granted from the Bradford Teaching Hospital NHS Foundation Trust.

5.3 General Format of the Primary Care

The availability, completeness, and level of detail in the data varies between systems and suppliers of the primary care GP systems. The providers supply healthcare systems and services to GP practices and associated organisations across the UK. Details of the data providers and the coding schema used are summarised in Table 5.1 below. The healthcare providers system for GP practices has adopted different coding classification as part of their underlying data schema.

The dataset used for this work comes from the TPP provider that uses SystmOne as the healthcare system in GP practices. It has been developed with continuous clinical input and has been operational as a single, remotely managed solution for 20 years. The system consists of all the functionality required, including the intuitive appointments, protocols, advanced clinical tools and reports for commissioning and monitoring, as well as fully integrated document management.

Table 5. 1 Summary of the Healthcare System Providers

NHS	*GP Computer Supplier*	*GP System*	*Clinical Coding Classification*	*Prescription Coding Classification*
England	TPP	SystmOne	Clinical terms version 3 / CTV3 / Read v3	British National Formulary (BNF)
	Vision	Vision	Read version 2 / Read v2	• Read version 2 / Read v2 • Dictionary of Medicines and Devices (dm+d)
Scotland	EMIS / Vision	EMIS / Vision	Read version 2 / Read v2	BNF
Wales	EMIS / Vision	EMIS / Vision	Read version 2 / Read v2	Read version 2 / Read v2

The basic format of the dataset extracted from any of the GP practice healthcare systems usually contains variables that are considered the most important for epidemiological research. It includes codes for clinical events (e.g., history, diagnosis, lab test result, symptoms, procedure), variety of administrative codes (e.g., registration, referral to a specialist), and prescriptions (e.g., medications that are prescribed but not necessarily dispensed).

- Registration records
    - Patient ID, registration date, and date of removal from practice lists. Multiple registration records are available per person from most (but not all) the suppliers.
- Clinical events
    - Date and clinical core (Read version 2 or CTV3) for primary care events, such as consultations, diagnoses, history, symptoms, procedures, laboratory test results, and administrative information. Where available, value fields have been modified to remove potentially identifiable information.
- Prescriptions
    - Date, drug code (Read version 2, BNF (see section 1.5.1.1.1 for further details) and/or dm+d) and where available, drug name, and quantity for medicines or devices prescribed in primary care. Drug name and quantity will assist with the interpretation of drug code fields that have different levels of completeness.

5.3.1 Clinical Coding and Classification System

Read code is a code thesaurus of clinical terms used in primary care since 1984 (Primary *et al.*, 2005). There are two versions: version2 (Read v2) and version 3 (CTV3 or Read v3). Both provide a standard vocabulary for clinicians to record patient findings and procedures. Read v2 and CTV3, together with a UK Read code browser, are available via the NHS Digital Technology Reference Data Update Distribution (TRUD) website. Read v2 and CTV3 were last updated in April 2016 and April 2018, respectively. Both versions are now deprecated (as the Read Browser), and no further updates will occur. From April 2018, SNOMED CT was introduced into primary care in a phased approach, and it is intended by April 2020 that SNOMED CT will be fully incorporated across the wider NHS, including codes related to the prescriptions.

5.3.1.1 Read Code Version 2 and 3

Read code version 2 is a five-character which earlier version of Read codes (version 1) is hierarchical, like a family tree. The version 2 code set was recommended for use by the Joint Computing Group of the British Medical Association, Royal College of Practitioners, and the Primary Health Care Specialist Group. The 5-byte codes set offers around 100,00 terms and are in use in most primary care computer systems in the GP practices in the UK.

Read code version 3 or CTV3 is a concept-based coding system developed in 1994 that has over 200,000 over terms. The intention was to develop a terminology that could include specialist practice as well as a general practice. The structure of the CTV3 allows a directed acyclic graph to replace the traditional hierarchy, permits multiple such graphs if necessary, introduce qualifiers, embedding these in an information model to support analysis, introduces one-to-many mapping to external classifications where this is necessary, and maintains the traditions of a dynamic terminology that stresses the inclusion of natural clinical terms.

5.3.1.2 British National Formulary (BNF)

The BNF is the standard catalogue of approved UK medical products, dressings, and appliances. It is released as a reference guide in both the online and paper versions and includes information for over 70,000 items on, for example, dosage, side effects, and quality. Code lists are manually modified and can be downloaded from the NHS Business Services Authority (NHSBSA). The BNF does not cover all items recommended by the NHS in its regular form, and certain

items are mentioned in the appendices as opposed to structured chapters. The NHSBSA has developed pseudo-BNF codes and chapters (18-23) to tackle this, including dressings, other medications, products and equipment.

The format and detail of the BNF codes varied by supplier and researchers are advised to that care must be taken to correctly interpret them, particularly when data from multiple sources are combined. The full BNF presentation code, as provided by the NHSBSA for chapters 1-15 and 18-19, is fifteen characters in length. The first seven of the code represent the categories of the medication in the BNF list and the last 8 digits represent the medicinal form, product, strength and the link to the generic equivalent product. Table 5.2 shows an example for Tradorex, a medication for opioid pain available in tablets, capsules and other forms.

Table 5. 2 Example of detail of level for Tradorex

BNF Level	Relevant Characters In BNF Code	Example Code: Tradorec XL Tablets 300mg	Description
CHAPTER	1 & 2	**04**0702040BIACAM	Chapter 4, central nervous system
SECTION	3 & 4	04**07**02040BIACAM	Section 4, analgesics drug
PARAGRAPH	5 & 6	0407**02**040BIACAM	Paragraph 2, opiod analgesics
SUBPARAGRAPH	7	040702**0**40BIACAM	Subparagraph 0, means the BNF does not extend to the subparagraph level
CHEMICAL SUBSTANCE	8 & 9	0407020**40**BIACAM	Tramadol hydrochloride
PRODUCT NAME	10 & 11	040702040**BI**ACAM	Tradorec
FURTHER PRODUCT INFORMATION (E.G. CAPSULE, TABLET, LIQUID STRENGTH)	12 & 13	040702040BI**AC**AM	Tradorex XL_Tab 300mg. Some may put the strength as 50mg = AA, 750mg = AB and 1000mg = AC, however, the letter do not always refer to the same strength of the medicines or drug
EQUIVALENT PRODUCTS	14 & 15	040702040BIAC**AM**	(i) The 14th and 15th characters will be the same with the 12th and the 13th characters if the product is generic (ii) The 14th and the 15th character will be the same if the product is the brand name (if exists) (iii) Code 'A0' for the 14th and 15th character for the brand and a generic equivalent where the product does not exist

Chapters 20-23 follow a similar coding format and are eleven character in length. They relate to dressings and appliances, hence no information on chemical

substance and dose is necessary. BNF codes are provided in this data interim release for TPP, however, the formatting of these codes differs by source.

5.3.1.3 Dictionary of Medicines and Devices (dm+d)

The prescription data from Vision (England) contains dm+d codes (as well as Read v2 codes) to record medicines prescribed to patients. The dm+d dictionary has been developed for use throughout the NHS (primary and secondary care) to identify specific medicines and devices used in the treatment of patients and consists of a dictionary containing unique identifiers and associated text descriptions.

The dm+d model consists of five components:

- A virtual Therapeutic Moiety (VTM) – the substances intended for use in the treatment of a patient
- Virtual Medicinal Product (VMP) – the properties of one more AMPs
- Actual Medicinal Product (AMP) – a single dose unit of an actual product known to have been available from a specific supplier
- Virtual Medicinal Product Pack (VMPP) – the properties of one or more equivalent AMPPs
- Actual Medicinal Product Pack (AMPP) – the packaged product supplied for direct patient use

An example of the dm+d component structure for a packet containing 56 tablets of Yaltormin 500mg is shown in Table 5.3. Note that the generic name appears in the VTM, VMP and VMPP dm+d components while the brand name is used in the AMP and AMPP components.

Table 5. 3 List of medication for polypharmacy

DM+D CODE	DM+D COMPONENT	DESCRIPTION
109081006	VTM	Metformin
386047000	VMP	Metformin 500mg modified-release tablets
35547511000001101	AMP	Yaltormin SR 500mg tablets (Wockhardt UL Ltd)
8990611000001109	VMPP	Metformin 500mg modified-release tablets 56 tablets
35547911000001108	AMPP	Yaltormin SR 500mg tablets (Wockhardt UL Ltd) 56 tablets

5.4 Data Profiling

5.4.1 Records and Classes of Data

The dataset consists of 227,482 patients age 65 years and above identified during the year 2020 (the year when the extraction of records was made from the live HIS). The data ranges from the clinical event data, timestamped, medications

prescription, and healthcare professional health care who recorded the data into each patients' EHR on every visit. Table 5.4 gives an overview of the main classes of the data.

Table 5. 4 List of attributes in Bradford dataset

Table	*Description*	*List of attributes*
Descriptive	Demographics details of the patient	`Digest, Sex, Age, Practice, Postcode, Ethnicity, Religion and MaritalStatus`
Clinical	Clinical data such as symptoms, diagnosis, procedure and lab measurement	`Digest, ReadCodeID, ReadCodeDesc, EventDate, EventTime, ConsultationMethod, ProfessionalContact`
Frailty score	The electronic frailty score calculated after the data extraction based on the EHR of patients	`Digest, NumDeficits, EfiCalc, Period`
Frailty Code	Lookup table for all the CTV3 Read Code associated with the 35 frailty deficits	`ReadCodeID, ReadCodeDesc, CTV3Code, CTV3Text, Deficit`
Prescription	Administration records of intravenous medication and medication orders	`Digest, MedicationName, StartDate, EndDate, Dosage`
Intervention	The start date of intervention	`Digest, StartDate`

The patient identifier has been unidentified by replacing the NHS number to the `Digest` number to secure the confidentiality of patients using the structured ID alteration steps. The number is 20 in length consists of the combination of numbers and alphabets. The `Digest` number used as the unique key identifier that linked to all the attributes in the relational database, as shown in Figure 5.1 below.

The data reference model in Figure 5.1 shows how data are distributed across the data tables represented using the Entity Relationship Diagram (ERD). The `ClinicalData`, `SRPrimaryCareMedication` and `EfiCalcLog` used to describe the main patient journey in primary care. The dictionary table for CTV3 read code associated with frailty deficits is in the table `CodeListsForAnalysisLkp` which contains approximately 2,095 unique codes. The frailty score recorded in the database is referring to a weekly score where the increment for unique frailty deficit is identified throughout the week.

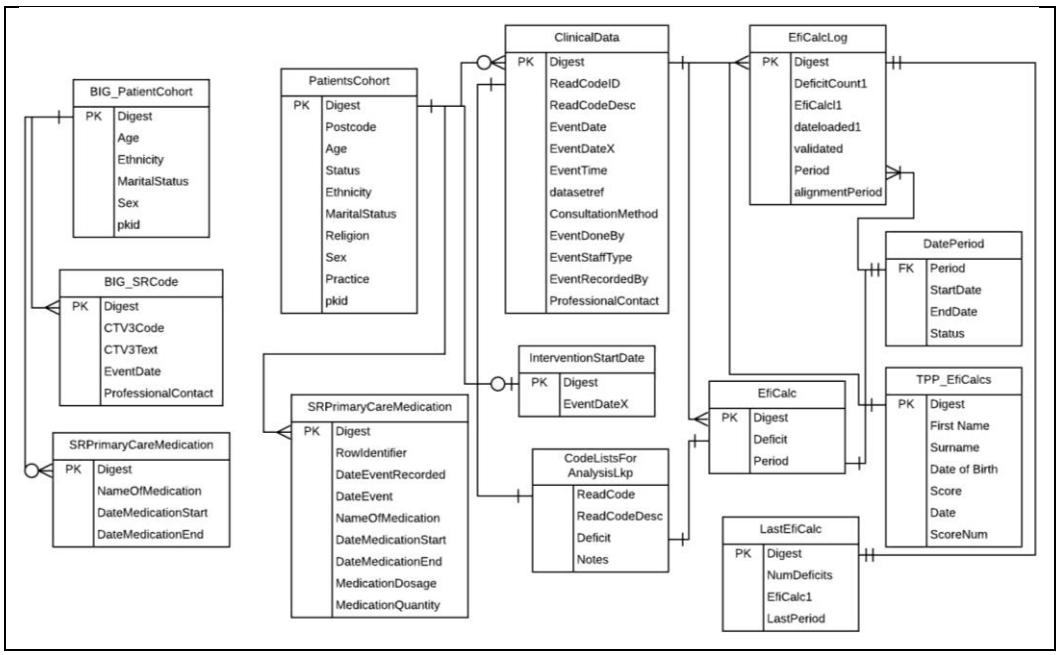


Figure 5. 1 The Entity-Relationship Diagram (ERD) of the Bradford Dataset.

5.4.2 The Sparsity and Density of the Data

The study is conducted on the retrospective dataset, where the data were gathered as early as the year 1877. The cut point of study duration will be identified to minimise the bias of the result. The sparsity and density dimension of the data across the timeline of the year plotted to determine the cut point for the study duration. The figure plotted as the dotted chart produced by process mining tool, ProM.

The input data accepted by the ProM must be in the form of an event log, where it consists if user ID (patient identifier/`DigestNumber`), date or timestamp, and the activities. The dotted chart displays the value of two variables from the dataset to identify how the patient's records change over time. The variables are the unique date each patient had in their record. It represents the number of the visit or any activity initiated during that time. The changes can be observed from the distribution of the data in each patients' record.

The significant gap between dot indicates high sparsity and less density in the year before 1900 as compared to the dot after that year. The next plot starts on the year 1900 until 2018 shows two different significant features: there is high sparsity before the year 1990 and high density after 1991; one straight empty line appeared during the year 1990. Figure 5.2 shows the first 1000 unique date for each patient, where the x-axis represents each unique date of a patient in the record and y-axis is the time in year measurement. The year 2003 start to show

consistent distribution along the year, and less significant gap and denser data observed towards the end of the data recorded.

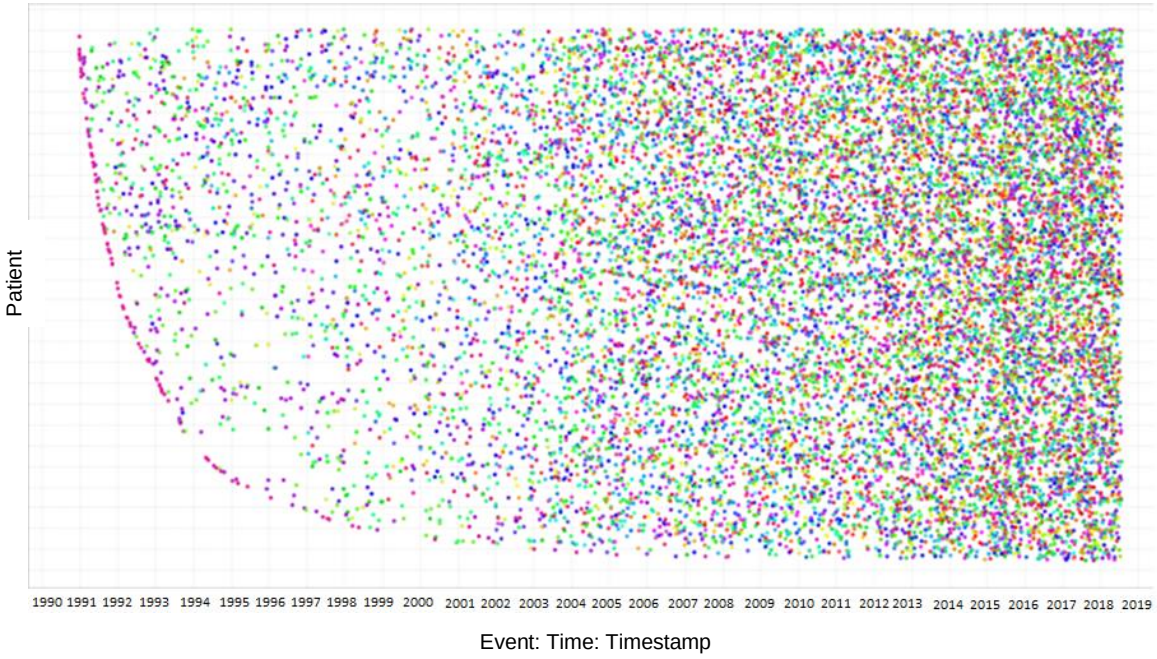


Figure 5. 2 The dotted chart for visualising the sparsity and density of the data recorded

The distribution of the total unique date in each patient recorded a gradually increased from the year 1991 until 2017. However, there is a 38% drop in the year 2017 to 2018 due to the incomplete record for that year. Hence, the total study duration chosen is 15 years, between the start of the year 2003 until the end of the year 2017.

5.4.2.1 Time Concept Drift

The main feature in the data where the process happened is normally assumed to be in a stable state (Bose, Aalst and Pechenizkiy, 2011). This type of secondary data usually being collected or recorded mainly not because of research purpose. The differences of the process observed at a certain period and this is where the time concept drift is rooted. The evolution of data over time is crucial to identify especially in pattern mining, machine learning, data mining, data streams, recommender systems and data retrieval to provide with the most accurate and precise in-process supporting of operational improvement.

In this task, it uses quite a similar approach as (Christian *et al.*, 2008) where the difference is we observe the event log projection over several years in the dotted chart and segregate them to focus into a smaller window of time to identify the changes. As stated in the (Bose and Aalst, 2013) focus on studying the feature

to monitor and identify when these characteristics changes are the start point in this task. This step is to identify the characteristics of the traces based on the dependencies or the relationship between activities. The aspect of the changes divided into three categories; control flow or behaviour perspective, data perspective and resource perspectives.

5.5 Data Quality Inspection

The 'real world' administrative, routinely collected data (such as this interim GP data release) have enormous potential to support research with far-reaching benefits to human health. By their nature analysing and interpreting these data within the context of health, research requires careful consideration of their content, structure, and crucially, an in-depth understanding that they were collected for an entirely different purpose. Hence, the quality of data extracted from this source is very low for analysis and requires a comprehensive step of processing.

The dataset is assessed with data quality assessment demonstrated by (Kahn *et al.*, 2016). The three categories of data quality include (1) completeness of the value presents in the data, (2) conformance of the value with the expectation values and (3) plausibility to check whether the data can be trusted.

5.5.1 Data Quality Category I: Completeness

The first data quality category is to check for the completeness of the dataset. It was done following the minimum requirement for process mining analysis as part of the verification step which user id, activity, and timestamp. The additional requirement may be included if present such as the resources.

Completeness determines the set of frequencies of data presents in the dataset on its own without referencing to its value. This step aims to identify the table which holds the complete patient identifier as there are several tables that may contain redundant values and incomplete identifier. The patient identifier defines as patient Digest number with additional patient information such as age, gender, and ethnicity. Acknowledge the missing and incomplete information contain in certain table would help us determining the data completeness with high quality to be included in our work.

Table 5. 5 Summary of data requirement attributes checking

Table Name	# row	Distinct Case ID	Missing Case ID	Missing Activity	Missing Timestamp	Missing Resource
PatientsCohort	13,020	13,020	0	-	-	-
ClinicalData	5,384,712	5,118	0	0	9,964 (0.19%)	0
SRPrimaryCareMedication	5,407,049	12,069	0	0	33 (0.06%)	11,336 (0.21%)
EfiCalcLog	4,588,173	4,781	0	0	-	-
EfiCalc	4,588,173	4,781	0	0	-	-
LastEfiCalc	4,194	4,194	0	0	-	0
TPP_EfiCalcs	486	486	0	0	-	-
InterventionStartDate	4	4	0	0	0	-
DatePeriod	1,486	-	-	0	0	-
CodeListsForAnalysisLkp	2,143	-	-	0	-	-
BIG_PatientCohort	235,830	235,830	0	0	-	-
BIG_PrimaryCareMedication	112,526,455	204,054	0	0	0	-
BIG_SRCode	139,718,196	227,482	0	0	0	0

The number and percentage of missing value for each table is assessed and summarised in Table 5.5. Each table was assessed following the requirement for process mining analysis represented by the last four columns in the Table 5.6. The next step is identification of the attributes value that consists of user id, activity, timestamp and additional attributes resource in each table of the database. Table 5.6 shows the summary of the data attributes checking of the dataset. In this dataset, the case ID is the `Digest` number which is a unique ID anonymised from the patient ID. The attribute activity for this dataset can be represented by either `Deficit`, `ReadCodeID`, `ReadCodeDesc`, and `NameOfMedication,` which are readily available. Each attributes activity indicates the level of granularity of the events for the analysis.

Table 5. 6 Summary of data requirement attributes checking

Table Name	Case ID	Activity	Timestamp	Resources
CodeListsForAnalysisLkp	No	*ReadCode, ReadCodeDesc, Deficit*	No	No
PatientsCohort	*Digest*	No	No	No
ClinicalData	*Digest*	*ReadCodeID, ReadCodeDesc*	*EventDate, EventDateX, EventTime*	*ConsultationMethod, EventDoneBy, EventStaffType, EventRecordedBy, ProfessionalRole*
SRPrimaryCareMedication	*Digest*	*NameOfMedication*	*DateEventRecordedBy , DateEvent, DateMedicationStart , DateMedicationEnd*	*MedicationDosage, MedicationQuantity*
EfiCalcLog	*Digest*	No	No	*DeficitCount1*
EfiCalc	*Digest*	*Deficit*	No	No
LastEfiCalc	*Digest*	*Deficit*	No	*NumDeficits*
TPP_EfiCalcs	*Digest*	No	No	No
InterventionStartDate	*Digest*	No	*EventDateX*	No
DatePeriod	No	No	*StartDate, EndDate*	No
BIG_PatientCohort	*Digest*	No	No	No
BIG_PrimaryCareMedication	*Digest*	*NameOfMedication*	*DateMedicationStart , DateMedicationEnd*	No
BIG_SRCode	*Digest*	*CTV3Code, CTV3Text*	*EventDate*	*ProfessionalRole*

For example, the *Deficit* attribute is the higher-level events of the *ReadCodeID*. It is a collection of code associated with frailty, which will be explained in a later section.

Timestamp attribute is recorded with different dimensions; attributes *EventDate*, *EventDateX*, *DateEvent*, *DateMedicationStart*, *DatemedicationEnd*, *StartDate,* and *EndDate* having date only and *EventTime* stored time entities with hours and minutes. We could get the complete timestamp (date, hours, and minutes) from the combination of attributes *EventDateX* and *EventTime*. Analysis requires additional resources attributes that can use the table *ClinicalData* and *SRPrimaryCareMedication*. For resource attributes, the higher-level event is the ProfessionalContact grouped from the attributes *EventStaffType*, *EventDoneBy*, and *EventRecordedBy* will be explained in the later section.

5.5.2 Data Quality Category II: Conformance

The second data quality category is conformance focuses on the representational agreement of the data with their formatting and relational structure. The first subcategories of conformance data quality are to determine whether the value conform to its establish set of format and standard. The data attributes involve is gender value stored in the *PatientsCohort* and *BIG_PatientsCohort* tables. It found that all gender values in *PatientsCohort* table has the character value of ‘M’ (48.3%) represents male and ‘F’ (51.7%) represent the female. However, the *BIG_PatientsCohort* tables has ‘M’ (41.45%), ‘F’ (58.46%) and additional two values of ‘N’ (0.07%) represent the unknown and ‘O’ (0.004%) as other.

The second subcategories of conformance are relational investigation of the data elements are accordance with the relational structure constraints of the database is done. The step involves identifying the number of patient record links to another table. The main patient identifier table are *PatientsCohort* and *BIG_PatientsCohort* tables which will be linked into the other two events tables containing the medical records: *ClinicalData* and *BIG_SRCode,* and prescriptions records: *SRPrimaryCareMedication* and *BIG_SRPrimaryCareMedication* summarised in Table 5.7. It presents the maximum number of patient identifiers match from the joined tables.

Table 5. 7 The relational conformance data quality presents the maximum number of patients that have records when linkage is done for both datasets.

Table Linkage	# patient
PatientsCohort	13,020
PatientsCohort + ClinicalData	5,118
PatientsCohort + SRPrimaryCareMedication	12,069
BIG_PatientCohort	235,830
BIG_PatientCohort + BIG_SRCode	227,482
BIG_PatientCohort + BIG_PrimaryCareMedication	204,054

5.5.3 Data Quality Category III: Plausibility

The last data quality category is plausibility which focuses on the trustworthiness of the data and determines by the variable's values. It divided into two subcategories: uniqueness and atemporal. The uniqueness aims to determine no duplication occur on attributes value that need to have exclusive value to represent itself. The Read codes of each deficit associated with frailty is examined to ensure no redundancy. Table 5.8 summarise the frailty deficit which codes has redundancy. A domain expert is involved to confirm the duplicated codes to resides in only one deficit for validation purpose. For example, a Read code *vascular parkinsonism* is shared between PT and PVD. After the validation step is taken the code vascular parkinsonism is placed in the PVD deficit only.

Table 5. 8 The redundant Read codes identification for plausibility uniqueness data quality. The second column represent the deficit whose codes was duplicated.

		Codes after Plausibility Step					
		Diabetes	MCP	PT	PVD	SU	Total
Codes Before Plausibility Step	CKD	14					14
	Diabetes					1	1
	PT		1		1		2
	PVD	4				1	5
	VI	6					6
					Total of Redundancy codes		**28**

The next subcategories are atemporal plausibility focuses on determining agreement between attributes values and the common knowledge. The step was done by checking the context of the code agrees with the gender value of each patient. For example, male patient will not be expected to have any codes related to pregnancy and no codes related to prostate disease is expected to be in female patient records. No disagreement was found between the patient gender and the context of codes recorded.

5.6 Event Log Extraction

The event log extraction for the dataset was done in several steps starting from creating initial event logs and performing transformation process as part of our data preparation and lastly the merging operation. The overview of the event log creation is presents in the Figure 5.3 to illustrate our approach.

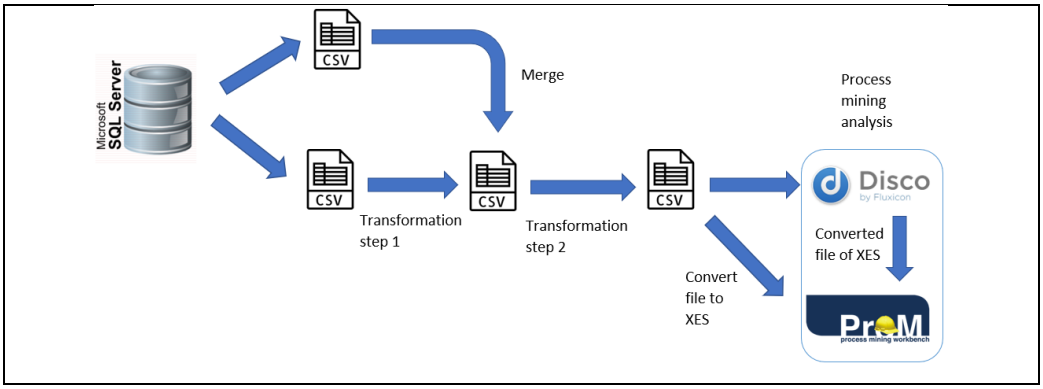


Figure 5. 3 The overview of the event log extraction flow

The raw dataset which came from two GP practices is managed in the SQL Server Management, and the initial extraction of the event log was done using SQL queries. The event log will undergo two transformation steps and one merging operation before it is uploaded into the process mining tools. Should be noted here, the standard file format for process mining is in XES (eXtensible Event Stream), which could be converted either using Disco or directly using ProM. Detailed extraction steps are as follows:

1) **Create two sets of initial event log**

   The first event log formed by the combination of descriptive, clinical, and frailty code tables; for the second event log, two tables of descriptive and prescription are joined where the sequence of the events is based on the temporal order. The reason for the separate extraction of the initial event log is to produce simplified input logs for the transformation steps. An example of SQL queries for the first set of an initial log is provided in Figure 5.4.

```
WITH satu AS (
SELECT cd.digest,  cd.ReadCodeID,
  cd.[EventDateX] as Timestamp
  FROM  [proj_NikChecker2].[dbo].[tblClinicalData] as cd )
  , dua AS (
  SELECT DISTINCT l.Digest, l.ReadCodeID, a.Deficit, l.Timestamp
  FROM satu as l
  FULL OUTER JOIN [proj_NikChecker2].[dbo].[tblCodeListsForAnalysisLkp_NEW] as a
  ON l.[ReadCodeID] COLLATE SQL_Latin1_General_CP1_CS_AS = a.[ReadCode] COLLATE SQL_Latin1_General_CP1_CS_AS
)
  , tujuh AS (
  SELECT DISTINCT Digest, Timestamp, ISNULL (Deficit, 'Nondeficit') AS Type_Record
  FROM dua
  WHERE Timestamp BETWEEN '2003-01-01 00:00:00' AND '2017-12-31 00:00:00'
  AND Timestamp IS NOT NULL )
  , tiga AS (
  SELECT pc.pkid, t.Timestamp, t.Type_Record
  FROM tujuh as t
  INNER JOIN [proj_NikChecker2].[dbo].[tblPatientsCohort] as pc
  ON t.Digest = pc.Digest
  WHERE pc.Practice IN ('Saltaire', 'Bingley')
  AND Type_Record != 'Nondeficit' )
  SELECT pkid, Timestamp, COUNT(Type_Record) as deficit_count
  FROM tiga
  GROUP BY pkid, Timestamp
  ORDER BY deficit_count DESc
```

Figure 5. 4 SQL queries for the first set of event log

2) Transformation Step1: Point of Polypharmacy Identification

The second set of initial event log is required as the input file that contains attributes such as patient id, name of prescription, medication start date, medication end date, and medication code. The transformation step will follow the approach explained in the section 1.5.1.5 for identification of polypharmacy deficit using an open-source interactive data science web tool Jupyter. The output file for this step will have attributes; patient id, type of deficit (polypharmacy), name of the medication, code of medication, and timestamp.

3) Merging operation

Both the output files from the previous step and the first set of initial event logs will be merged to produce a total of 36 frailty deficits.

4) Transformation Step2: Calculation of Frailty Score

Once we have the full set of frailty deficit, the frailty index score is calculated at every clinical event present within the patient record. The re-calculation of the eFi score is conducted to have an overall score and frailty category for each event. A snippet of the code for the re-calculation of the eFi score is shown in Figure 5.5.

```
import csv

filename = "admissions_4above.csv"
exclude = ["nondeficit"]
total_deficit = 27

subjects = []
rows = []
with open(filename, newline='') as csvfile:
    spamreader = csv.reader(csvfile, quotechar='|')
    n = 0;
    for row in spamreader:
        if (n > 0):
            data = [row[0], row[1].replace('"', ""), row[2].replace('"', "")]
            rows.append(data)
            subjects.append(row[0])
        n += 1
    subjects = list(set(subjects))
    f = open("efi_4above.csv", "w")
    output = "subject_id,timestamp,deficit,eficount,efiscore\n"
    for subject in subjects:
        counted = []
        efi = 0
        for data in rows:
            x = []
            if(data[0] == subject):
                x.append(data[0])
                x.append(data[1])
                x.append(data[2])
                if((data[2] not in counted) and (data[2] not in exclude)):
                    counted.append(data[2])
                    efi += 1
                    x.append(efi)
                else:
                    x.append(efi)
                efiscore = round((efi/total_deficit), 3)
                output += str(data[0]) + "," + data[1] + "," + data[2] + ","
                    + str(efi) + "," + repr(efiscore) + "\n"
    f.write(output);
```

Figure 5. 5 The re-calculation code for eFi score

5.7 Data Preparation

In this section, we will discuss data transformation with a log enriching approach in detail. It includes the identification for point of polypharmacy and calculation of frailty score following the existing method by (Clegg *et al.*, 2016).

5.7.1 Polypharmacy

Polypharmacy is an expression used to describe a condition of simultaneously used of multiple different types of medication by an individual (Masnoon *et al.*, 2017) and commonly due to the multimorbidity (Nobili *et al.*, 2011) presence in the population. It used to explain both the good and bad situations of polypharmacy, which depends on the management of medications. Optimised medication usage where prescription made based on the best evidence is relevant polypharmacy whereas, the irrelevant polypharmacy when multiple of improper medications prescribed or the planned benefit is not aware (Nicholson and Stone, 2013). It also stated that inappropriate medications should be avoided as possible, and any evidence for choosing medication (e.g., patient preferences of taking such medication) that might lead to polypharmacy should be documented and improved.

The polypharmacy constitutes with the problem when the risk of adverse drug events and drug to drug interactions is prominent which is common in the elderly (Rieckert *et al.*, 2018) as they tend to develop more chronic diseases as age

increases. The consequences of polypharmacy comes with higher healthcare costs, medication non-adherence, decrease functional status and geriatric syndromes (Strehl, 2013). It is one of the thirty-six frailty deficits proposed by (Clegg *et al.*, 2016) that affect the healthy ageing of the elderly patient.

5.7.1.1 Terminology of Prescription Mapping

This part aims to assess the inclusiveness of prescription extracted from the dataset and will act as an input for the polypharmacy deficit identification step which will be described in the next section. The list has been made after taking consideration of clinician insight in determining medication that could cause potential harm especially to the elderly patients. The list consists of the first six of BNF code with 724 of unique code and will be referred as the `exclusion document`. The list has been excluded as it has no potential systemic effect on the body. Some example of prescriptions includes are for allergies, influenza, antivirals, antibacterial, most appliances and devices.

The prescription data extracted from the dataset was followed the dm+d prescriptions terminology. It contains the dm+d code with the patient identifier, the date prescriptions been dispensed, the start and end date of medicine consumption and the dosage of the prescription attached. The dm+d codes consist all the components with 13,693 of unique dm+d product descriptions of the prescriptions and will be referred as the `medication document`. The third documents which involve in the mapping steps is called the `mapping document` which can be found in website (https://www.nhsbsa.nhs.uk/prescription-data/understanding-our-data/bnf-snomed-mapping). This document represents mapped data between dm+d and the Master Data Replacement Drug database which also has a field showing the BNF code of prescriptions and SNOMED code as shown in Table 5.9. The document published in January 2020, which also contains information relating to November 2019.

Table 5. 9 List of components in the Mapping Document

COMPONENT	TOTAL
PRESENTATION / PACK LEVEL	2
VMP / VMPP / AMP / AMPP	4
BNF CODE	48,295
MDR: PRODUCT DESCRIPTION	48,038
SNOMED CODE	334,105
DM+D: PRODUCT DESCRIPTION	102,070
DM+D: PRODUCT AND PACK DESCRIPTION	106,076

The total number of medications included for polypharmacy is 4,762, which is based on the first six digits of the BNF code identified to be prevalence within the elderly. The prescription code is based on the BNF pharmaceutical reference book that contains a broad spectrum of information and advice on prescribing and pharmacology, along with the specific's facts and details about various medicines available on the UK NHS. It consists of fifteen different chapters where the selected medications prescribed constitute about 35% of the total number of medications record extracted.

5.7.1.2 The Point of Polypharmacy

In this part, we are assessing the list of medications relevant within the elderly to identify polypharmacy with an assumption that all patients are taking the medication prescribed to them. Polypharmacy is defined by consuming at least five different medications in a day (Dagli and Sharma, 2014). The list of medications prescribed for each patient will be assessed to determine the point of polypharmacy begins.

Formal definitions are provided in order to give more clarity in the approach used to identify polypharmacy with illustration in Figure 5.6 as a guide. The definitions follow:

Definition 1 (Medication event). A medication event consists of a medication name M, ($M_1, .., M_n$) with the length of the course to take the medication including the start date S, ($S_1, .., S_n$) and end date E, ($E_1, .., E_n$). The structure of the event may happen in sequence or parallel within the record.

Definition 2 (Point Pp). The first point of polypharmacy identified within the study duration in the form of date.

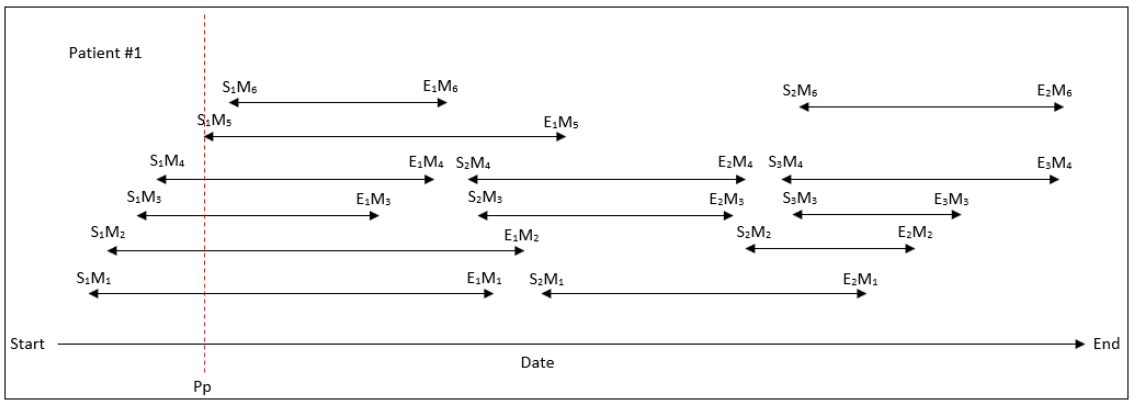


Figure 5. 6 The illustration for point of polypharmacy identification for a patient

The steps in identifying the point of polypharmacy are performed by accessing each patient record of different medications within the study duration. The point `Pp` is determined iteratively at the start of the record to locate a single date that intersects with at least five different medications. The first point identified is the `Pp`, which will be the date for the activity polypharmacy.

5.7.2 Electronic Frailty Index Score

The collection of the Read code to form deficits will be discussed in the pre-processing section 1.2.8.1.2 below. For example, if five deficits present out of 36, the frailty score would be 0.14, and it is in category mild. The complete list of the range of frailty score and its category is illustrated in Table 3.2 in Section 3.2.3.1.

Currently, the eFI feature has been implemented in the SystmOne and readily available within the reporting tab. A report of the eFI score and category is generated within the GP practice and can be downloaded in the CSV format. The eFI score and frailty category are not included in the clinical record as it is only generated for the selection of a population at a recent time. The re-identification of the eFI score and category is being carried out outside the SQL Server environment as the score is required for each visit for analysis.

Comparison checking of eFI score is performed between a generated report and re-identification score to ensure correctness and consistency of the score obtained. Overall, we obtained a similar eFI score from both reports.

5.8 Event Log Pre-processing (Data Transformation)

In this part, the data will be prepared from the raw data for analysis in the next chapter. It will undergo a process of cleaning and transformation to acquire event logs for process mining research. It is one of the crucial steps in the process mining, which has been mentioned in (Van der Aalst *et al.*, 2012) as the initial challenges in process mining analysis.

In this section, we perform several pre-processing approaches to provide a standard event log suitable for process mining analysis in healthcare. The standard event log should produce a process model in which the level of complexity is low. We implement two event log pre-processing steps to form a collective of high quality of low-level events in the log and dealing with the temporality issue.

5.8.1 Pre-processing I: Events Abstraction using Ontological Concept

The clinical data consists of low-level events that compromise with the quality of the process model later. Hence, the aggregating events step is conducted for determining the suitable event for frailty deficits and the professional roles involved in performing the tasks. The identification of such events was made with the help of the domain expert. The aggregating events and dealing with temporality issues steps in this section were done in the SQL Server database and Python Jupyter platform after the healthcare system's extraction.

5.8.1.1 Resource Events (Professional Role)

The first pre-processing of aggregating events is done within the resource events. The majority of the data quality issue in any healthcare application is the high granularity of low event events produced (Orfanidis, Bamidis and Eaglestone, 2004), which mainly affected by its usability, accessibility, and availability. In order to ensure optimum interoperability of the healthcare system, the integration between different systems being established. As a result, the patient records will include care received from various care team members within a different specialty.

Table 5. 10 The mapped staff type into a different type of professional role

STAFF TYPE	PROFESSIONAL ROLE
Therapists	Allied Healthcare Professional
Podiatrist	
Optician	
Physiotherapist	
Dietician	
Pharmacy Technician	
Doctor	GP
GP Associate	
Specialist Registrar	
Sessional GP	

These low-level events which share similar properties and have clear semantic meaning can be represented as ontological events. Within the clinical department, the staff holds the difference in job scope, position level, or specialty, which make up various types of staff. For example, a GP with different staff types had been providing treatment to the patient within different visits to the GP—some examples of mapped low-level events to the designated professional role illustrated in Table 5.10.

This approach had dramatically reduced the number of events and activities performed by health professionals. However, the number of variations in the process is still extremely high and require other processing methods.

5.8.1.2 Activity Events (Frailty Deficits)

The second pre-processing of aggregating events is done within the activity events. The activity events which consists of Read codes will undergo two stages of mapping of fine-grained using the ontological concept which are (1) mapped of codes into associated frailty deficit and (2) mapped of codes into clinical classes as illustrated in Figure 5.7 which will be further explained in next section.

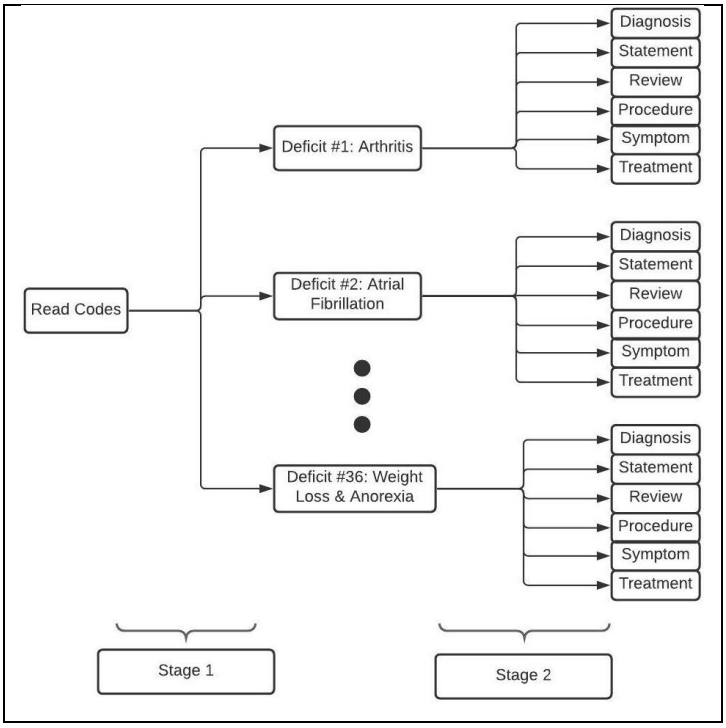


Figure 5. 7 The overview of the two stages of codes mapping

Stage 1: Mapped of Codes into Associated Frailty Deficit

The first stage of aggregating activity events is done by mapping Read codes into associated frailty deficit. The accumulation of molecular damage and unrepaired cellular will develop deficits as the cells age (Rockwood and Mitnitski, 2011). The accumulated deficit will present as age-related health issue such as diseases and disabilities.

As we have explained in section 1.2.7.3 determining the frailty index score following the previously described eFi (Clegg *et al.*, 2016) we identified total of 1,909 Read codes prevalence within the elderly with the help of domain expert. The code is mapped into 35 different frailty deficits as illustrated in Table 5.11.

Table 5. 11 The distribution of the number of codes associated in each frailty deficit

DEFICIT	# CODES	DEFICIT	# CODES
Activity Limitation	8	Ischaemic heart disease	60
Anaemia Haematinic Deficiency	192	Memory/Cognitive Problems	104
Arthritis	76	Mobility/Transport Problems	23
Atrial Fibrillation	22	Osteoporosis	49
Cerebrovascular Disease	75	Parkinsonism and Tremor	31
Chronic Kidney Disease	38	Peptic Ulcer	102
Diabetes	140	Peripheral Vascular Disease	22
Dizziness	32	Respiratory Disease	223
Dyspnoea	13	Requirement for Care	13
Fall	12	Sleep Disturbance	11
Foot Problems	9	Skin Ulcer	97
Fragility Fractures	120	Social Vulnerability	17
Heart Failure	47	Thyroid Disease	42
Hearing Impairment	36	Urinary Incontinence	33
Housebound	13	Urinary System Disease	30
Heart Valve Disease	3	Visual Impairment	136
Hypertension	39	Weight Loss & Anorexia	14
Hypotension/Syncope	27		
Sub-Total	**902**	**Sub-Total**	**1,007**

The total number of events that resulted in this step is reduced to 90% of the original number of events. Although the number of activities and events is reduced significantly, the number of variations of the process is still high and appears not giving a significant impact on this step. The high percentage of reduction should not be a worry as to the loss of information for the process. The reasons being are the collective of the event selected are prevalence within the elderly patient, which is the focus of this work to highlight the significant pattern of the health process.

Stage 2: Mapped into Clinical Classes

The second stage of aggregating activity events is mapped of Read codes of each deficit into clinical classes. The codes belong to a particular concept should share the similar properties and have clear semantic meaning such as diagnosis, symptoms, clinical procedures, clinical statement, healthcare review and treatment. An example of the code descriptions mapped to the healthcare concepts is illustrated in the Table 5.12 with two deficits; diabetes and respiratory disease. For most of the analysis we focus on the diagnosis type of code with a total 592 codes.

Table 5. 12 Mapping of code description into healthcare concept using ontological concept approach

Deficit	Read Code	Description of code	Healthcare Concept
Diabetes	XE10H	Diabetes mellitus with neurological manifestation	Diagnosis
	XaJOj	O/E – left eye preproliferative diabetic retinopathy	Procedure
	XaJLa	Diabetic retinopathy 12-month review	Review
	F4200	Background diabetic retinopathy	Statement
Respiratory Disease	H3…	Chronic obstructive lung disease	Diagnosis
	XaIQT	Chronic obstructive pulmonary disease monitoring	Review
	XaIUt	COPD self-management plan given	Statement
	1713.	Productive cough-clear sputum	Symptom

5.8.2 Pre-processing II: Dealing with Temporality Issue

This method aims to avoid getting misunderstandings with the process model pattern. Temporal inconsistency is an approach taken to overcome the issue and will be discussed in a later section.

5.8.2.1 Temporal Inconsistency

The sequence of events of the event log is ordered based on their timestamp. Hence, the consistency of the time format for the timestamp used in the event log will ensure the right pattern of the process model. First, the recognition of the timestamp format provided in the dataset was done (see Table 1.3). The only table that records the highest granularity of timestamp (hour and minutes) is `ClinicalData`. The table contains the clinical event and deficits with full timestamp format recorded with date, hour, and minutes. However, for polypharmacy deficit, the timestamp recorded with date only. Combining it and other clinical events (in `ClinicalData`) would result in polypharmacy appearing to precede the clinical events if they are occurring on the same date. To address this issue and maintain the consistency of the temporality of the events, we select only 'Date' without the timestamp detail.

5.9 Cohort Selection and Data Characterisation

The cohort for the study was undergo three stages of selection as illustrated in Figure 5.8. The initial extraction from the live HIS patient record yield 227,482 patients aged 65 years and above during the extraction year at 2020. During the first stage of selection, only patient who has record within the study duration. The

number of patients is further reduced to only include elderly patient at the start of study duration. The eligibility of elderly patients is defined by the age starting from 65 years old and above at the start of study duration following the work (Ritt *et al.*, 2015; Clegg *et al.*, 2016; Lansbury *et al.*, 2017). The selection criteria at stage two excluded almost half of the initial extracted patient. The final stage stratified the elderly into frail and non-frail elderly by the frailty identification tool as explained in section 1.7.2.

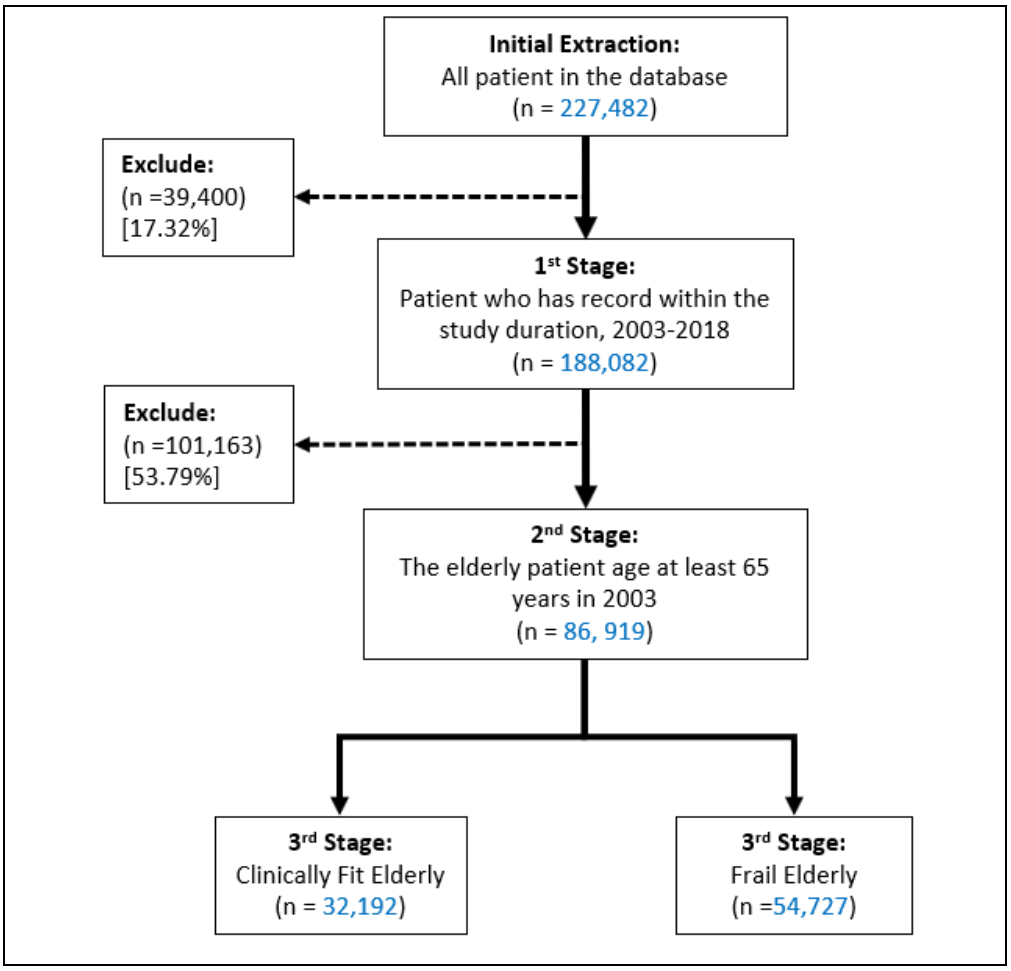


Figure 5. 8 The stages of cohort selection

The distribution of elderly patient with respect to their gender is presented in Table 5.13. The clinically fit elderly patient constitutes about 37.0% from the study population, follows by the 33.0% from the mild category, 21.0% from moderate and 9.0% from the severe category. The female patient is higher by 15.9% from the male patient, which support that women had higher life expectancy compared to men. Although the number of elderly in the clinically fit category is higher than the frail elderly with difference only 4.1%, it is quite contrast from the distribution percentage in the previous work in developing the eFi score (Clegg *et al.*, 2016).

Table 5. 13 The distribution of the elderly patient in the selected cohort

Category	Female	Male	N	U	Total
Clinically Fit	17,845	14,306	40	1	32,192
Mild	16,638	11,992	21	3	28,654
Moderate	11,139	7,128	7	0	18,274
Severe	5,204	2,614	0	0	7,799
Total	**50,817**	**36,030**	**68**	**4**	**86,919**

5.10 Conclusion

This chapter discussed the initial steps taken in our analysis, which is the acquisition of the event log. The dataset was from the EHR of elderly patients extracted from the primary care setting. The raw dataset has undergone the initial selection of the study duration based on the sparsity and density of the data distribution across the years. Two data preparation steps were used to identify the polypharmacy and calculation of the eFi scores. The next two event log pre-processing steps, which are aggregating events and dealing with the temporality issue of the log was applied. All steps had prepared the event log for our analysis in frail elderly in the next chapter. The progressive technological advancement of healthcare system makes the routinely collected data in primary care a rich resource for research. However, using a real healthcare records impose a big challenge for any research work, especially in the consistency of the healthcare record storing format. Low granularity of the event which provides only dates present a limitation. The issue become a problem when analysing the administration events as the order of event within same day could not be refined.

Chapter 6

Modelling Frailty Progression

The first case study was presented in Chapter 4 involving the analysis using MIMIC-III dataset as the preliminary work. This chapter demonstrates the experiments conducted using the Connected Bradford dataset as the second case studies. Works in the chapter focused mainly in determining whether process mining able to model frailty progression using the clinical data from primary care setting. Section 6.1 will explore the background of work in deriving the experiments in this chapter. Section 6.2, 6.3 and 6.4 present the experiments conducted to analyse frailty progression.

6.1 Background

This section explores the overview of Connected Bradford dataset used in the experiments. It includes data diversity and representativeness. Next, background related to the work of frailty progression will be presented.

6.1.1 Overview Dataset for Experiments

Connected Bradford dataset was introduced in detail in the Chapter 5. The dataset comprises elderly people aged 65 years and above who have at least one year record within duration of 2003 – 2018. The representative of the cohort is selected based on stratification steps of average accumulation of frailty deficit to ensure cohort not included patient whose already passed the frailty threshold (Bartosch, McGuigan and Akesson, 2018). The richness and readily available of EHR data provided advantages over the complete variables in measuring frailty within the primary care setting. Polypharmacy is an example of variable that easily accessible in the nature of EHR which fit with the polypharmacy definition compared to MIMIC-III dataset (Mortazavi *et al.*, 2016; Masnoon *et al.*, 2017). Data abstraction step based on clinical concepts is introduced to achieve the aim of experiment one. The rest of the experiments follow similar extraction steps.

6.1.2 Frailty Progression

Frailty is a geriatric state common in elderly (Clegg *et al.*, 2013). It is a syndrome differs from the comorbidity and disability which results in lowering body functional reserve (Lally and Crome, 2007). The dramatic deteriorations in body structure and function could be observed physically is in musculoskeletal systems influencing mobility, balance, and capability to live without any support or assistance (Gielen *et al.*, 2012; Wilson *et al.*, 2017). It portrays by declining response to increase internal and external bad events which leads to increase risk to hospitalisation, dependency, institutionalisation and even death (Fried *et al.*, 2004).

The immense impacts frailty given not only to the elderly people themselves, but to the family, society and healthcare system has attracted attention of researchers and clinicians. The increasing change demographic of elderly population worldwide became additional indicator to understand frailty and its clinical implications to achieve successful and healthy ageing (Kojima, Liljas and Iliffe, 2019). The process of frailty is dynamic where the transition of frailty is likely to occur from non-frail state to frail and more frail state over time (Gill *et al.*, 2006; Lang, Michel and Zekry, 2009). However, the fluctuating nature of frailty mostly happened in a single direction from non-frail to frail state and ending in the frail state (Espinoza, Jung and Hazuda, 2012; Setiati *et al.*, 2019).

The trajectories or progression of frailty is varied among individuals and affecting by many factors (Welstead *et al.*, 2020). It was reported in some literatures study regarding the difference trajectories based on 1) clusters of frail elderly patient identified as three groups; “developing frailty”, “maintaining frailty” and at “high risk of frailty”, 2) four trajectories of “relatively stable, mild, moderate and severely frail”, 3) two trajectories based on age groups of 60-69, 70-79 and 80-89 (Hsu and Chang, 2015; Chamberlain *et al.*, 2016; Verghese *et al.*, 2021). The frailty trajectories studied were based on five-years interval assessment and annual assessment from the baseline year with closed-monitor patient cohort. Although, a study found to utilise the EHR based on monthly assessment, network like trajectories representation based on ordered diagnosis is still lack (Stow, Matthews and Hanratty, 2018). Apart from that, the association between frailty progression and common diseases in primary care related to frailty is still under studied.

6.2 Experiment 3: Frailty trajectories pattern within frailty categories

The MIMIC-III dataset been used previously to examine the suitability of combining process mining and eFi with routinely collected record in analysing the variability of frailty trajectories within categories. Although the previous experiment evidently stated that the combination of process mining and eFi is possible, the statistically significance trajectories of diagnosis is still not achieved. Thus, the analysis in this experiment will only focused on diagnosis-based frailty deficit. The experiment is based on the combination method following literatures in analysing disease trajectories using statistical approach and process mining (Jensen *et al.*, 2014; Westergaard *et al.*, 2019; Kusuma *et al.*, 2020).

6.2.1 Stage I: Planning

This experiment aims to analyse the differences in significant frailty trajectories pattern between frailty categories. The primary research question was "*Are combination of process mining and statistical steps possible in exploring significant pattern of diagnosis-based frailty trajectories?*" with the following research question of "*What is the significant trajectories within frailty categories*"? The contribution of this experiment is improvement of the previous approach taken in experiment 3 by providing a systematic method in modelling frailty trajectories to analyse the trajectories of diagnosis variability within categories. The research question of this experiment addressed (RQ1), (RQ2), (RQ3), (RQ4) and (RQ5) of the research. The (RQ1) focusses on both data (statistical) and process mining approach in determining the significant trajectories using systematic (RQ4) approach of data extraction. The best visualization is made viable (RQ2) in displaying the significant frailty trajectories as it progresses from clinically fit to severe categories.

6.2.2 Stage II: Extraction

The extraction of dataset is initially obtained following in Section 5.6. Three fundamental requirements for process mining analysis for dataset are 1) case ID, a unique ID to represent each cases or patient in the data. The **case ID** is Digest ID used as the patient identification in the database of Connected Bradford dataset. The second requirement is 2) **activity** names is derived from the frailty deficit and lastly 3) is **timestamp** to determine the sequence of activities. The

event log is created by combining all records of activities with the respective timestamp identified from the `ClinicalData` table.

6.2.3 Stage III: Data Transformation and Loading

The event log is then undergoing event log processing steps. The first step is 1) log enriching where category of frail patient is determined as explained in Section 5.7 which is part of data preparation step to either be in clinically fit, mild, moderate, or severe categories based on the final cumulative frailty deficits within the study duration.

6.2.3.1 Selection of diagnosis-based Frailty Deficit

The next step is 2) log filtering which incorporating the selection of diagnosis-based frailty deficit. The frailty deficits consist of diagnosis, disability, symptoms, clinical statements, procedures, and test results. Diagnosis-based frailty deficits are defined as the clinical evidence of having in any health condition associated with frailty recognised by the healthcare professionals. Five out of 36 frailty deficits; housebound, social vulnerability, mobility and transport problems, requirement for care, and activity limitation are all identified as diagnosis, a recognition for a certain health condition occurs in elderly. Table 6.1 illustrates two out of five frailty deficits code description mapping: activity limitation and social vulnerability.

Table 6. 1 Mapping of code description to clinical concept

Deficit	Read Code	Description of code	Clinical Concept
Activity limitation	9EB5.	Disability	Diagnosis
	Y3502	Allowance / DLA applied for	Diagnosis
	Y3501	Already receiving attendance allowance / DLA	Diagnosis
Social Vulnerability	13M1.	Death of spouse	Diagnosis
	XaJvD	Does not have a carer	Diagnosis
	ZV603	Person living alone	Diagnosis

The second stage of log filtering is applied as only the first occurrence of diagnosis-based frailty deficits are considered. Finally, artificial ‘Start’ and ‘End’ events are added into each case. The ‘Start’ event is added as the first event and ‘End’ event is added as the final event. The event addition is required because of the limitation of plugin in ProM which cannot deal with multiple start and end event in the log. The approach in identifying the frailty trajectories is following the method discussed in Chapter 3 in Section 3.3.1. It starts with quantifying disease association, testing the directionality of pair of deficits and creating an event log.

6.2.4 Stage IV: Mining and Analysis

This stage includes the process mining and process analytics from the output of the previous stages. The event log of each category after performing steps of identifying the frailty trajectories are mild (80.01%), moderate (86.45%) and severe (98.33%) resulted from the processed event log as illustrated in Table 6.2. The number of cases and events are reduced significantly as shown in Table 6.2 after undergoing two main of screening stages. The first stage of screening is where the steps of identifying frailty trajectories involves measuring the strength of association between deficit diagnosis and the second screening is applying the variant filter in Disco. The variant filter aims to reduce the complexity of frailty trajectories model by reducing parts of distinct case behaviour in the log to obtain the significant frailty trajectories. It was done by selecting cases that share sequence of deficit diagnosis by at least 2 cases. It worth noticing that the highest percentage of case variation is severe (90%) which is expectable as severe constitutes the largest number of unique frailty deficits.

Table 6. 2 Details of transformation step and final event logs.

Description	Mild	Moderate	Severe
	Processed Event Logs		
# patient	28,653	18,274	7,799
# pair of diagnosis	1,257	630	630
	Filter #1: Measure RR		
# pair of diagnosis	508	508	508
	Filter #2: Calculate CI		
# pair of diagnosis	493	495	484
	Filter #3: Co-occurrence with Pareto Principle		
# pair of diagnosis	100	85	154
	Event Logs		
# patient	22,941	15,797	7,669
# activities	25	25	26
# total variants	6,612	7,192	7,098
Variant percentage*	28.82%	45.53%	92.54%
	Variant Filtration		
# patient	18,035	10,060	801
# total variant	1,706	1,455	230
Variant percentage*	21.23%	14.46%	28.71%

6.2.4.1 Diagnosis Association based Frailty Deficit Co-occurrence

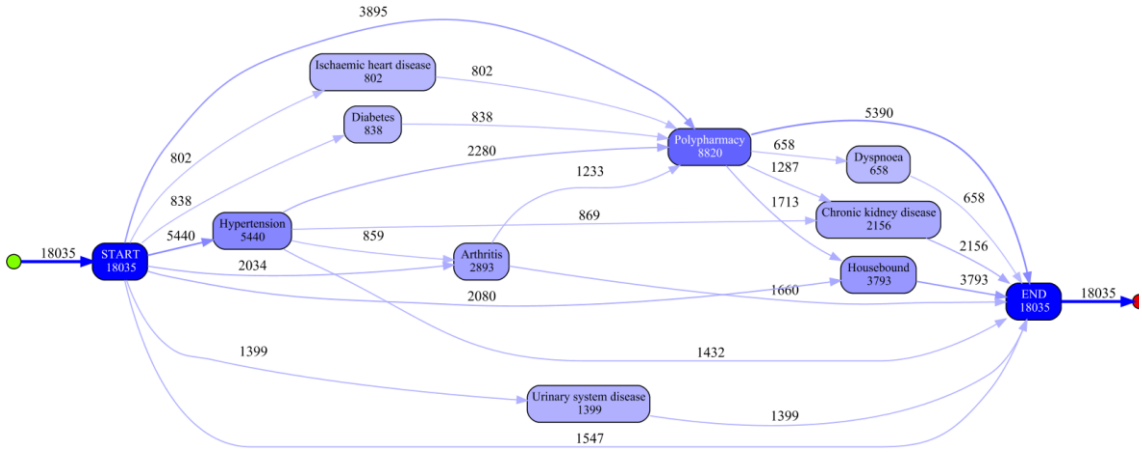
The pair of deficit diagnosis will only include if at least 179 patients had in mild, 218 in moderate and 93 in severe. The co-occurrence of minimum count of diagnosis pair was following the pareto principal method. The goal was to reduce the number diagnosis to achieve the 80% threshold where the significant trajectories come from the 20% pair of diagnosis with 80% frequency of

happening. A total of 339 directional pairs (referred as the deficit diagnosis co-occurrence) which had an elevated relative risk and preferred statistical direction were identified.

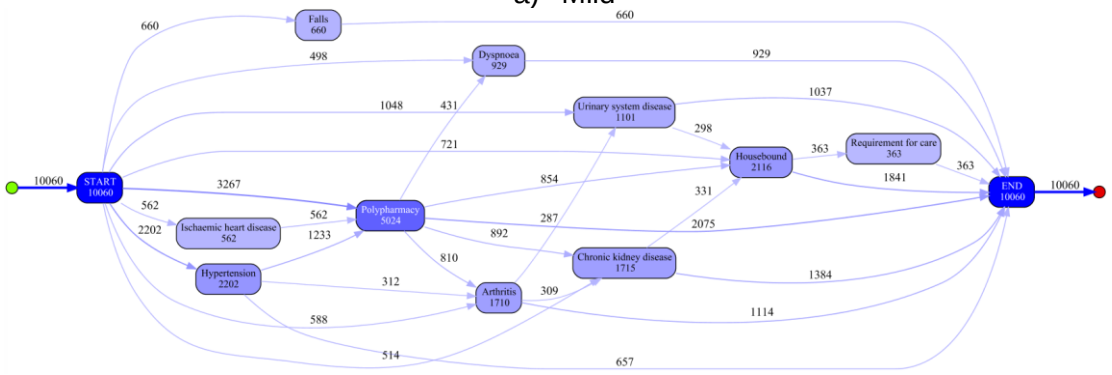
The dominant directional pair was found to be between hypertension → polypharmacy in all categories. The arrow indicates the direction of the pair. The pair represents in 2,455 mild patients (10.7%), in moderate about 1,759 patients (11.1%) and 776 severe patients (10.1%) in each sub-population respectively. The next significant deficit diagnosis following the dominant pair and shared by two categories is Polypharmacy → Housebound encountered in 1,825 mild patients (8.0%), in 1,183 moderate patient (7.5%) and Arthritis → Polypharmacy involved in 409 (5.3%) severe patients. The third ranked pair of deficit diagnosis are varying among the three categories where mild and moderate have similar pair of deficit diagnosis with different direction; Polypharmacy → Housebound occurred in 1,825 (8.0%) in mild and Housebound → Polypharmacy occurred in 1,183 (7.5%) in moderate while pair of Diabetes → Polypharmacy found in 409 (5.3%) in severe patient.

6.2.4.2 Diagnosis based Frailty Deficit Trajectories

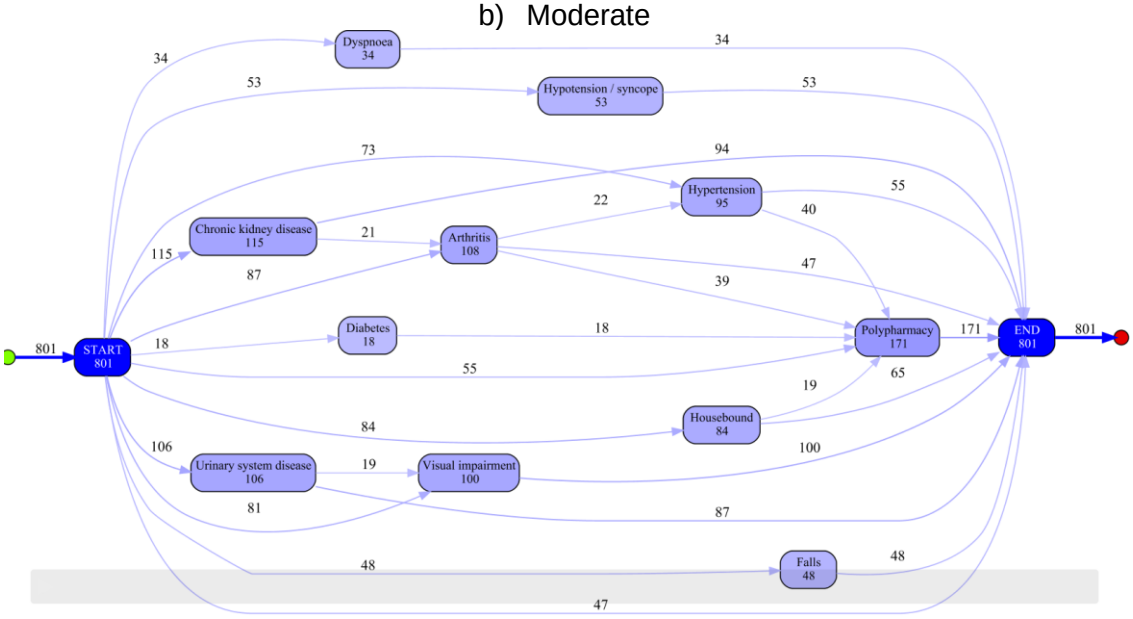
The event logs from each category are carefully assessed to only includes pair of deficit diagnosis resulted from the *RR*, *CI*, and Pareto Principal method with 100 pairs in mild, 85 pairs in moderate and 154 in severe. The creation of frailty trajectories as part of process discovery was done using the plugin Directly Followed Visual Miner (DFvM). The plugin is chosen as it produced the most readable workflow models compared to others plugin. The process models produced are set to visualise 50% frequent activities and 50% frequent paths in all models.



a) Mild



b) Moderate



c) Severe

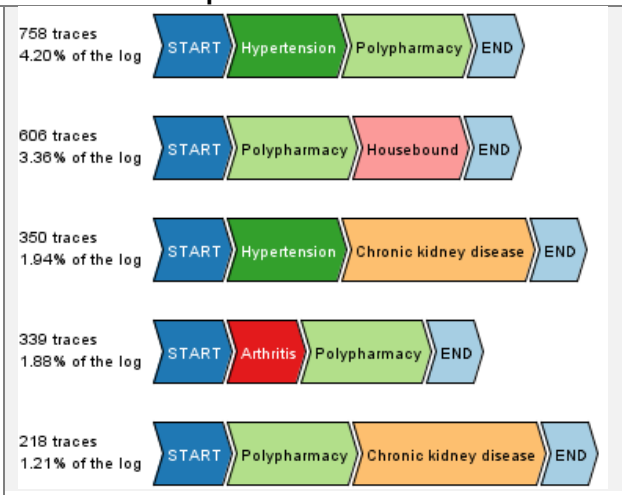
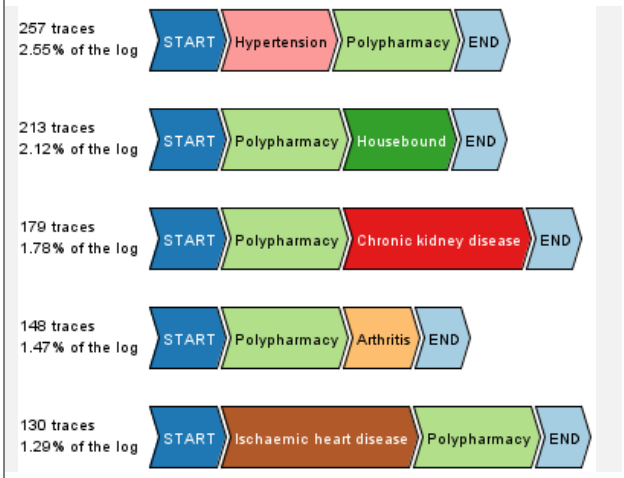
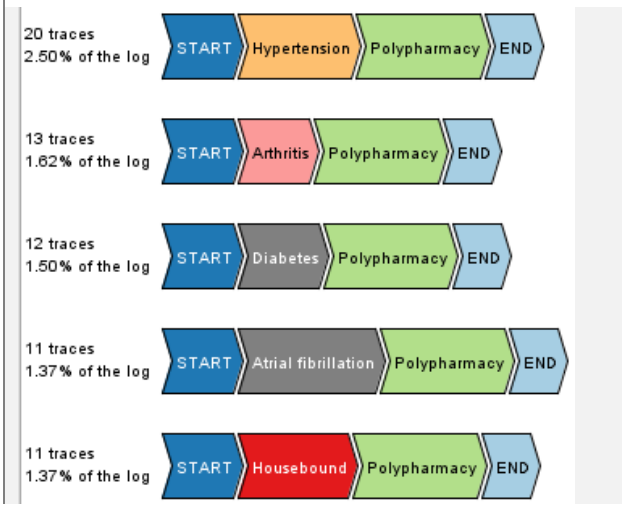
Figure 6. 1 The frailty trajectories for a) mild, b) moderate and c) severe using DFvM in ProM

Figure 6.1 represents three frailty trajectories of mild, moderate, and severe. The polypharmacy is the most frequent diagnosis found in all category; 9,005 (18.7%) in mild, 5,119 (17.9%) in moderate and 211 (10.9%) in severe. As it can be

observed in mild, Hypertension, Arthritis, Ischaemic heart disease and Diabetes come first before Polypharmacy. The deficit diagnosis identified to occurred after Polypharmacy is Chronic kidney disease, Housebound and Dyspnoea. The directly followed trajectories dominating the a) mild model can be found starting from Hypertension to Polypharmacy and Housebound. While in b) moderate, Hypertension and Ischaemic heart disease are found to come before polypharmacy, whereas Housebound, Chronic Kidney Disease, Arthritis, Urinary system disease and Dyspnoea come after Polypharmacy. The directly followed trajectories which significant in the model apart from Hypertension to Polypharmacy is from Polypharmacy to Chronic kidney disease. The directly followed deficit diagnosis can be observed to occur between Chronic kidney disease to Arthritis and finally to Polypharmacy in c) severe patient.

The exploration of trace variant of event log is analysed using the trace variant feature in ProM tool as shown in Table 6.3. Trace variant is the differences of diagnosis sequence and the feature allow to examine the dominant behaviour of diagnosis which is followed by at least (1.2%) in all category. The most common trace variant in all category is Hypertension followed by Polypharmacy, mild happened about (n = 758, 4.2%), moderate (n = 257, 2.6%) and severe (n = 20, 2.5%). Polypharmacy is common in severe where Arthritis, Diabetes, Atrial fibrillation and Housebound found to happen before. While in moderate Polypharmacy mostly happened before with total percentage of (5.4%) in three out of five trace variants. It differs in mild as the proportion of trace variant associated with Polypharmacy has similar total percentage between Polypharmacy occurred first then followed by either Housebound and Chronic kidney disease and Hypertension and Arthritis followed by Polypharmacy.

Table 6. 3 The top five trace variant which is common in each category

Category	Top five Trace Variant
Mild	
Moderate	
Severe	

Further analysis was done on the process metric of each event log. The overview statistics of event logs is illustrated in Table 6.3. Severe have the least number of cases in the event log, the mean case duration is the shortest compared to mild and moderate. The average events per cases is the least in severe. The composition of the severe contains the least percentage of cases (10%), events (4%) and trace variants (3%) which following the condition of variant filter to only select trace that shared by at least two cases. Mild contains cases (78%), events (66%), trace variant (25%) while moderate contains cases (63%), events (46%) and trace variants (20%) from the event log with has highly associated pair of deficit diagnosis. Although smallest number of cases, event and trace variants observed in severe, the variant percentage is highest as compared to other categories. This suggest that high variety of pair of deficit diagnosis in severe patients due to maximum numbers of diagnosis in severe is 36 as compared to moderate 12 and mild 8.

Table 6. 3 Overview statistics of frailty trajectories models of mild, moderate, and severe

Category	# Cases	# Events	Events per cases	Case duration
Mild	18,035	48,047	Min 2, Mean 3 and Max 6	Median: 1 Year and 9 Months Mean: 2 Years and 10 Months
Moderate	10,060	28,498	Min 2, Mean 3 and Max 6	Median: 1 Year 7 Months and 3 Weeks Mean: 2 Years 8 Months and 3 Weeks
Severe	801	1,945	Min 2, Mean 2 and Max 6	Median: 0 Year and 10 Months Mean: 1 Year and 10 Months

6.2.5 Stage V: Evaluation

Evaluation of this experiment has been done by analysing the suitability of combining process mining with statistical approach in exploring the significant pattern of diagnosis-based frailty trajectories. The statistical approach was implemented to reduce the complexity of the trajectories. The approach was able to present the high associated diagnosis related to frailty within the frail population. The main reason DFvM plugin is chosen over commercialise and user-friendly model, Disco tool is because it will intentionally create the low association of diagnosis. It happened as the process of piercing the two directional pairs of high association between four different diagnoses of a, b, c and d, where $a_{T1} \rightarrow b_{T2}$ and $c_{T3} \rightarrow d_{T4}$ occurred in temporal order of T1→T2→T3→T4 in a similar case will create the link between b to c which is not significant pair.

The comparison of frailty trajectories between category was done by looking into the network structure of the model with the similar model settings. It found some

interesting result based on the significant pattern that Polypharmacy is highly associated in all categories. Polypharmacy found to be the prominent deficit diagnosis associated with several diagnosis: hypertension, housebound and arthritis. The association found as a result of the necessity to treat multiple disease (comorbidity) which is common in elderly (Nwadiugwu, 2020). However further analysis is required to determine at which frailty state that Polypharmacy is most common.

The general clinical comment of the results in this experiment is the combination of process mining and statistical approach is potentially valuable in generating the prevalence frailty trajectories within the elderly. The trace variant visualization apart from the process model is useful in facilitating the healthcare professional to understand the workflow of frailty. It is particularly helpful in presenting the variation of common trajectories within different frailty categories of mild, moderate, and severe. However, the downside is in the cohort selection strategy as it was done by stratifying the patient according to their final frailty score. This may generate a rather less simplified version of process model which illustrate the unbalanced frailty trajectories within different categories. Cohort stratification at each frailty category which will examines a fixed number of deficit progression is suggested for next analysis.

6.3 Experiment 4: Analysis of Frailty Progression

This experiment was done to explore the frailty progression over time. Frailty is known as an age-related state which makes individual more vulnerable to injury, hospitalisation, disability, and mortality. Even though they are commonly associated with natural aging process, it is not an unavoidable part of it. Hence, studying frailty is critical in understanding why some elderly becomes weak at faster pace than others. In hope with better knowledge, it helps to better identify and treat who are at most risk of deterioration. The aim of the experiment was to investigate how the course of frailty evolves based on the clinical records associated with frailty.

This section investigates the frailty progression analysis over time within elderly. This experiment follows the methodology explained in Chapter 3. The progression of frailty is analysed within four different frailty stages: not frail to mild, mild to moderate, moderate to severe and in final stage severe.

6.3.1 Stage I: Planning

The first stage in this work was done by understanding the dataset to plan the experiment accordingly. The scope of this experiment is the analysis of the elderly clinical record within the primary care in Bradford County region. Four research questions were addressed to study the frailty progression which are:

1) *Can frailty progression be determined based on the clinical records over time?*

2) *Is it possible to use process mining to assess the variability of frailty progression within the point of concern?*

3) *What are the significant patterns of frailty trajectories within cohort with point of concern in different elderly patient classification?*

4) *Where do the shortest and longest transition time occurred within the cohort with point of concern?*

The first research question was clarified by analysing the events (clinical records) associated with frailty to identify different type of frailty progressions. The inquisitive point of concern linked to frailty progression is 'high' progression and was tackled in research question two. It was explored by selecting elderly patients accompanying the point of concern and all their events. These address the (RQ3), (RQ4) and (RQ5) of the research in facilitating the possibility to analyse frailty progression in relation to eFi using the cut off points reported in literature. Moreover, the third and last research questions were addressed to determine the significant pattern of frailty trajectories associated with point of concern and its duration which relate to the (RQ1) and (RQ2) of the research. The hypothesis was that clinical record is possible to be used to study frailty progression within elderly and variation of significant pattern of frailty trajectories could be identified within different stages using process mining and data-based approach. The experiment used Jupyter Python for data processing and transformation, Tableau for visualisation, Disco and ProM tools for mining analysis.

6.3.2 Stage II: Extraction

The extraction of dataset for this experiment was done in two steps, first step involves acquiring a large cohort of dataset and second stage is applying the exclusion steps to carefully remove patient who is unfit for the study. The first step is to include patient aged 65 years above and have record within study duration of 2003 – 2018 as followed in the previous experiment. Next, exclusion

criteria are applied to exclude (i) patient who has less than one year of record associated with frailty deficit (ii) patient aged 85 years above and in clinically fit or early stage of frailty, mild and (iii) patient with average accumulation of frailty deficit more than three in a year.

The justification for the second exclusion criteria is to ensure that initial point of clinical record within study duration is selected at the early stage of frailty as mortality record of patient is unavailable within the study. The next exclusion criteria are based on past literature which found that most elderly patient has about 6%-7% increment per year (Bartosch, McGuigan and Akesson, 2018). The increment provides as a threshold for patient selection. A patient was found to be in an advanced state of frailty and progressing at a faster rate once he/she passed the threshold limit. At this stage, about 58,145 patients are consider suitable for the experiment.

6.3.3 Stage III: Data Transformation and Loading

The third stage of the experiment is the data transformation and loading to create an event log for analysis with process mining. Two steps are involved which are determining the rate of frailty progression and defining the type of frailty progression which will be discussed in detail in the next section. The approach of frailty progression analysis is explained in Chapter 3 in Section 3.3.1. It begins with quantifying the rate of frailty progression, discretization of the frailty progression rate values and evaluating the frailty deficits association.

6.3.4 Stage IV: Mining and Analysis

The discretisation approach has resulted in the splitting points of frailty progression values into three separate intervals as shown in Table 6.4. Low progression is the dominant definition of frailty progression covered by 87.3% of events experienced by all patients (100%) in cohort. Then it followed by medium progression occurred about 8.9% experienced by 41.1% patients in cohort. It worth noting that, the number of deficit increment in all definition of interval is similar but not for the number of days of interval for deficit the deficit to increase. No overlapping of duration was identified between the definition intervals. Even though, high progression constitutes the lowest events (3.8%) of all frailty progression definition, it would be considered as point of concern in frailty progression analysis.

Table 6. 4 Result after implementing discretization approach into the dataset

Definition Frailty Progression	Rate of Progression Interval	# Events (%)	# Patients (%)	Deficit Increment	Duration (days)
High	0.00429 – 0.08000	15,768 (3.8)	12,065 (20.8)	1	1 – 7
				2	1 – 14
				3	1 – 18
Medium	0.00010 – 0.00428	37,123 (8.9)	23,895 (41.1)	1	8 – 30
				2	15 – 60
				3	19 – 80
Low	0.00001 – 0.00099	359,003 (87.3)	58,145 (100.0)	1	31 – 5,624*a
				2	61 – 5,444*b
				3	81 – 4,687*c

*The maximum duration of days in Section_1 of *a occurred in about 15.4 years, *b 15 years and *c in 13 years

The progression of frailty was further examined by creating four event logs based on the frailty stages which starts from (Stage I) clinically fit or not frail, (Stage II) mild, (Stage III) moderate, to (Stage IV) severe as shown in Table 6.5. The purpose of this approach was to answer research question number two *is it possible to use process mining to assess the variability of frailty progression within the point of concern?* As the frailty stages course from less rigorous states (fit and mild) to rigorous (moderate) and to more advanced state (severe), the number of patients were expected to be 58,145 in fit stage and lesser in the next stages. The percentage of patients within the respective frailty stages were based on an individual patient who had at least one definition of each frailty progression. The percentage of events and patients were calculated based on the total number of events and patient at each frailty stages.

Table 6. 5 Details of dissection approach of patient records at each frailty stages

Frailty Stages	Definition Frailty Progression	# Events (%)	# Patient (%)
(I) Not Frail > Mild	High	5,535 (3.4)	4,882 (8.4)
	Medium	14,055 (8.6)	11,937 (20.5)
	Low	**142,991 (88.0)**	**57,473 (98.8)**
Total in Stage I		162,581	58,145
(II) Mild > Moderate	High	6,374 (4.0)	5,402 (10.9)
	Medium	14,410 (8.9)	11,796 (23.8)
	Low	**140,705 (87.1)**	**48,655 (98.1)**
Total in Stage II		161,489	49,595
(III) Moderate > Severe	High	2,978 (4.2)	2,527 (10.0)
	Medium	6,604 (9.4)	5,455 (21.7)
	Low	**60,599 (86.4)**	**24,238 (96.3)**
Total in Stage III		70,181	25,168
(IV) Severe	High	881 (4.1)	727 (9.7)

	Medium	2,054 (9.7)	1,631 (21.8)
	Low	**18,349 (86.2)**	**7,163 (95.5)**
Total in Stage IV		21,284	7,499

The dominant definition of frailty progression among all frailty stages is low, where the percentage of events are more than 80.0% and occurred in more than 90.0% of patients. The trend is decreasing throughout the frailty stages from (Stage I) fit to (Stage IV) severe by as little as 0.2% for percentage of events and 0.7% for percentage of patients. However, the opposite trend is observed for medium frailty progression. Medium frailty progression has an increasing trend observed in both event and patients with percentage increment between frailty stages as little as 0.3% and minimum percentage of patient increment at only 0.1%. Apart from that, overall trend for frailty progression is increasing from the (Stage I) fit until (Stage III) moderate stages, before declining slightly in (Stage IV) severe. The difference could be observed in 0.1% of events and 0.3% of patients.

Although, both proportion of events and patients for high frailty progression is the least within all frailty stages, it is point of concern in the frailty progression study as it is providing many intrigued questions such as what happened at point of high frailty progression, what is the impact and factor lead to that point? in many frailty studies. Along with this justification, the next step in this experiment is to investigate patients recorded with high frailty progression as the point of concern in this study.

Further analysis was conducted focusing on the point of concern in frailty progression. Elderly patients with the absence of high progression at any frailty stages were filtered out. The selected elderly patients were grouped based on age-specified cohort for the classification of elderly adults as youngest-old age between (65 – 74), middle-old age between (75 – 84) years and oldest-old aged 85 years and above (Lee *et al.*, 2018). The analysis is based on elderly age-specified classification since the frailty states worsened as elderly aged which affected about 10% in the adult aged 50 – 64 and 43.7% in elderly aged 65 years and above (Fogg *et al.*, 2022). The classifications of elderly adults will facilitate in determining the variation of frailty progression by establishing the age gap appropriately. The classified elderly patients were further stratified based on gender to allow the investigation of differences between male and female patients. Table 6.6 presents the descriptive statistics of sub-cohorts of elderly patients which includes the events of frailty definition distribution and the trace duration in years.

Six sub-cohorts representing male and female in each elderly age-based groups shown in Table 6.6 below. It is observed that female is the leading in both number of patient and event in all age classification, with the highest number in the 75 – 84 group. Although, low frailty definition still constitutes the highest number of events, events of high frailty definition were followed before event of low frailty definition in all age group and gender. The highest event of high frailty definition is dominating by the female in all age group 65 – 74 (3,363, 6.1%), 75 – 84 (4,938, 8.9%) and 85+ (1,243, 11.5%). The trace duration calculated between the start and end of dissected events within each frailty stages. The median trace duration shows that male had the shorter duration compared to female in all age groups age group 65 – 74 (9.8 years), 75 – 84 (7.2 years) with the shortest in the 85+ age group 5.3 years.

Table 6. 6 Descriptive statistics of sub-cohort of patient with high point of frailty progression

Age Range	**65 - 74**		**75 - 84**		**85 ++**	
Gender	Male	Female	Male	Female	Male	Female
# Patients	2,194	2,697	2,377	3,673	296	839
# Events	21,693	28,473	23,000	38,958	3,123	8,948
	# Events of Frailty Definition					
High	2,696	3,363	3,069	4,938	406	1,243
Medium	1,564	2,104	2,018	3,245	358	931
Low	25,095	20,207	15,358	26,795	2,034	5,848
	Trace Duration (in years)					
Average	9.4	10.2	7.5	8.3	5.7	6.4
Median	9.8	11.1	7.2	8.2	5.3	5.9
Min	1.0	1.0	1.0	1.0	1.2	1.0
Max	15.5	15.6	15.5	15.7	14.2	15.2
25%	6.2	7.3	4.2	5.1	3.9	4.1
75%	12.9	13.5	10.3	11.4	7.4	8.1

6.3.4.1 Significant Pattern of Frailty Trajectories

Individual frequent deficits contributing to the high frailty progression and their highest percentage out of all deficits presents at respective age group and gender are polypharmacy (14%), respiratory disease (5.2%), Housebound (7.1%), Urinary system disease (6.9%), Fall (5.7%), Diabetes (5.0%), Hypertension (5.3%), Atrial fibrillation (6.2%), and anaemia haematinic deficiency (5.2%). It revealed that polypharmacy is leading with the most frequent deficit contributing to the high frailty progression in all age groups and gender. It follows by housebound deficit presented in both gender in 75 – 84 group, female in 65 – 74 group and male in 85++ group. Meanwhile, urinary system disease is the third frequent deficits associated with high frailty progression. This deficit is commonly presented in female of 85++ and 65 - 74 group, while male in 75 – 84 and 85++ groups.

The variability of the frailty progression involving the point of concern is further explored by examining each frailty stages and their associated deficits contributing to the high progression. It shows the pattern of individual deficit that is commonly presents within different gender and age groups. It illustrates that individual deficits mostly presented at the early stage of frailty; stage I (Fit > Mild) and stage II (Mild > Moderate). Furthermore, polypharmacy deficit is dominating the stage I (Fit > Mild) and stage II (Mild > Moderate) with range of 1.4% - 20% based on the total count of deficits at each frailty stage. Surprisingly, hypertension is the second leading deficits (with the exception in male of age group 85++) mostly in female in all age groups.

The exploration of deficits at each frailty stages produces an intuitive knowledge of frailty progression albeit not directly answering the research questions. However, to discover the variability of frailty progression, identifying its significant pattern of trajectories is crucial. Hence, top five (5) pattern of significant frailty trajectories within frailty stages for every age sub-cohorts in male and female were shown in Table 6.7 (a) for age sub-cohort (65 – 74), (b) age sub-cohort (75 – 84) and (c) age sub-cohort (85++). The pattern of frailty trajectories is the common trace variant of each respective event logs.

Table 6. 7 Pattern of significant frailty trajectories

(a) Elderly aged 65 -74 with median is measured in months

# Traces (%)	Trace Variant	Median (months)	# Traces (%)	Trace Variant	Median (months)
Male			Female		
Frailty Stage I: Fit > Mild					
Median: 3.7 [# unique trajectories: 73]			Median: 5.1 [# unique trajectories: 68]		
89 (17.2)	Ischaemic heart disease > Polypharmacy	1.5	110 (15.6)	Polypharmacy > Diabetes	3.2
65 (12.6)	Diabetes > Visual impairment	3.9	107 (15.2)	Polypharmacy > Respiratory disease	4.0
35 (6.8)	Dyspnoea > Respiratory disease	0.2	60 (8.5)	Diabetes > Visual impairment	4.9
26 (5.0)	Diabetes > Pulmonary valve disease	6.7	55 (97.8)	Hypertension > Chronic kidney disease	27.2
25 (4.8)	Diabetes > Foot problem	4.8	40 (5.7)	Respiratory disease > Dyspnoea	5.8
Frailty Stage II: Mild > Moderate					
Median: 5.8 [# unique trajectories: 111]			Median: 5.0 [# unique trajectories: 61]		
18 (3.7)	Polypharmacy > Fall	10.2	28 (6.7)	Osteoporosis > Polypharmacy	1.7
17 (3.5)	Dyspnoea > Heart failure	1.0	25 (6.0)	Fragility fracture > Osteoporosis	1.2
15 (3.1)	Dyspnoea > Respiratory disease	1.1	25 (6.0)	Dyspnoea > Respiratory disease	0.7
15 (3.1)	Respiratory disease > Visual impairment	7.7	23 (5.5)	Fall > Housebound	3.0
14 (2.9)	Dyspnoea > Atrial fibrillation	1.6	22 (5.3)	Chronic kidney disease > Urinary system disease	11.5
Frailty Stage III: Moderate > Severe					
Median: 6.7 [# unique trajectories: 111]			Median: 8.2 [# unique trajectories: 162]		
11 (3.1)	Urinary system disease > Chronic kidney disease	6.9	22 (3.8)	Heart failure > Chronic kidney disease	1.6
11 (3.1)	Housebound > Requirement for care	6.9	16 (2.8)	Housebound > Requirement for care	3.5
10 (2.9)	Memory cognitive problem > Requirement for care	17.2	15 (2.6)	Housebound > Memory cognitive problem	5.3
9 (2.6)	Dyspnoea > Heart failure	0.2	13 (2.3)	Memory cognitive problem > Requirement for care	4.0
9 (2.6)	Housebound > Fall	2.6	12 (2.1)	Urinary system disease > Memory cognitive problem	5.4
Frailty Stage IV: Severe					
Median: 11.2 [# unique trajectories: 115]			Median: 13.3 [# unique trajectories: 231]		
7 (3.8)	Memory cognitive problem > Requirement for care	17.1	7 (2.0)	Housebound > Heart failure	9.7
6 (3.2)	Requirement for care > Housebound	4.9	7 (2.0)	Fall > Housebound	8.5
4 (2.2)	Memory cognitive problem > Housebound	8.3	6 (1.7)	Fragility fracture > Housebound	0.6
4 (2.2)	Visual impairment > Housebound	8.7	5 (1.4)	Fragility fracture > Osteoporosis	12.6
4 (2.2)	Urinary system disease	6.0	5 (1.4)	Anaemia & haematinic deficiency > Atrial fibrillation	4.6

(b) Elderly aged 75 - 84 with median is measured in months

# Traces (%)	Trace Variant	Median (months)	# Traces (%)	Trace Variant	Median (months)
Male			Female		
Frailty Stage I: Fit > Mild					
Median: 4.6 [# Unique trajectories: 83]			Median: 5.1 [# unique trajectories: 132]		
73 (11.3)	Ischaemic heart disease > Polypharmacy	1.6	109 (10.7)	Polypharmacy > Ischaemic heart disease	3.4
62 (9.6)	Diabetes > Visual impairment	6.7	89 (8.8)	Polypharmacy > Respiratory disease	3.6
58 (9.0)	Respiratory disease > Polypharmacy	4.4	63 (6.2)	Polypharmacy > Diabetes	1.8
46 (7.1)	Hypertension > Chronic kidney disease	12.6	54 (5.3)	Hypertension > Chronic kidney disease	23.9
28 (4.4)	Hypertension > Cerebrovascular disease	8.9	48 (4.7)	Dyspnoea > Respiratory disease	0.8
Frailty Stage II: Mild > Moderate					
Median: 5.8 [# unique trajectories: 121]			Median: 4.4 [# unique trajectories: 102]		
51 (8.3)	Polypharmacy > Housebound	9.4	41 (6.2)	Visual impairment > Chronic kidney disease	9.2
25 (4.1)	Dyspnoea > Heart failure	1.3	37 (5.6)	Polypharmacy > Requirement for care	3.3
19 (3.1)	Visual impairment > Chronic kidney disease	5.3	31 (4.7)	Osteoporosis > Polypharmacy	0.7
18 (2.9)	Visual impairment > Fall	19.0	29 (4.4)	Visual impairment > Hearing impairment	8.3
17 (2.8)	Fall > Fragility fracture	9.7	25 (3.8)	Hypotension / syncope > Fall	0.9
Frailty Stage III: Moderate > Severe					
Median: 6.7 [# unique trajectories: 124]			Median: 6.0 [# unique trajectories: 184]		
13 (3.4)	Memory cognitive problem > Requirement for care	8.8	28 (3.9)	Housebound > Requirement for care	10.3
12 (3.2)	Fall > Hypotension / syncope	2.3	23 (3.2)	Fragility fracture > Osteoporosis	2.3
10 (2.6)	Hearing impairment > Chronic kidney disease	9.0	21 (2.9)	Housebound > Memory cognitive problem	6.1
9 (2.4)	Fall > Visual impairment	9.4	20 (2.8)	Fall > Fragility fracture	1.4
9 (2.4)	Fall > Fragility fracture	3.0	18 (2.5)	Memory cognitive problem > Requirement for care	8.9
Frailty Stage IV: Severe					
Median: 7.1 [# unique trajectories: 165]			Median: 12.0 [# unique trajectories: 259]		
6 (2.8)	Housebound > Memory cognitive problem	3.0	7 (1.6)	Housebound > Requirement for care	12.7
3 (1.4)	Chronic kidney disease > Peripheral vascular disease	16.5	7 (1.6)	Fragility fracture > Memory cognitive problem	15.4
3 (1.4)	Memory cognitive problem > Hearing impairment	4.2	7 (1.6)	Fragility fracture > Osteoporosis	4.7
3 (1.4)	Housebound > Mobility transport problem	9.6	6 (1.4)	Urinary system disease > Memory cognitive problem	13.5
3 (1.4)	Housebound > Requirement for care	16.7	6 (1.4)	Anaemia & haematinic deficiency > Housebound	12.5

(c) Elderly aged 85++ with median is measured in months

# Traces (%)	Trace Variant	Median (months)	# Traces (%)	Trace Variant	Median (months)
Male			Female		
Frailty Stage I: Fit > Mild					
Median: 4.1 [# unique trajectories: 30]			Median: 4.2 [# unique trajectories: 68]		
7 (7.6)	Heart failure > Polypharmacy	3.6	17 (7.1)	Polypharmacy > Arthritis	2.5
6 (6.5)	Hypertension > Hearing impairment	9.2	14 (5.9)	Polypharmacy > Respiratory disease	3.8
6 (6.5)	Polypharmacy > Respiratory disease	3.9	12 (5.0)	Polypharmacy > Diabetes	4.3
6 (6.5)	Fall > Housebound	3.6	12 (5.0)	Hypertension > Chronic kidney disease	14.7
6 (6.5)	Arthritis > Housebound	9.3	9 (3.8)	Hear failure > Polypharmacy	1.6
Frailty Stage II: Mild > Moderate					
Median: 4.1 [# unique trajectories: 24]			Median: 3.7 [# unique trajectories: 45]		
6 (10.7)	Memory cognitive problem > Polypharmacy	2.2	21 (11.2)	Housebound > Fall	6.2
6 (10.7)	Respiratory disease > Polypharmacy	2.2	15 (7.9)	Hypertension > Polypharmacy	8.7
5 (8.9)	Visual impairment > Housebound	3.2	10 (4.8)	Fall > Urinary system disease	2.1
5 (8.9)	Anaemia & haematinic deficiency > Urinary system disease	8.7	9 (4.8)	Heart failure > Polypharmacy	0.9
4 (7.1)	Arthritis > Requirement for care	7.9	8 (4.3)	Heart failure > Chronic kidney disease	3.3
Frailty Stage III: Moderate > Severe					
Median: 8.3 [# unique trajectories: 5]			Median: 5.7 [# unique trajectories: 30]		
4 (50)	Urinary system disease > Housebound	19.9	5 (8.6)	Chronic kidney disease > Dyspnoea	10.6
1 (12.5)	Mobility transport problem > Osteoporosis	0.7	4 (6.9)	Fragility fracture > Requirement for care	11.9
1 (12.5)	Dyspnoea > Heart valve disease	4.8	4 (6.9)	Fall > Urinary system disease	2.1
1 (12.5)	Requirement for care > Memory cognitive problem	28.2	4 (6.9)	Heart failure > Polypharmacy	0.8
1 (12.5)	Heart valve disease > Visual impairment	1.2	4 (6.9)	HF > CKD	7.1
Frailty Stage IV: Severe					
Median: 6.0 [# unique trajectories: 19]			Median: 10.7 [# unique trajectories: 11]		
1 (5.3)	Housebound > Mobility transport problem	7.2	4 (23.5)	Requirement for care > Memory cognitive problem	13.4
1 (5.3)	Fall > Hypotension	0.0	3 (17.1)	Chronic kidney disease > Fragility fracture	13.1
1 (5.3)	Housebound > Chronic kidney disease	14.5	1 (5.9)	Urinary system disease > Housebound	20.1
1 (5.3)	Skin ulcer > Parkinsonism & tremor	4.9	1 (5.9	Urinary system disease > Housebound > Social vulnerability	14.1
1 (5.3)	Ischaemic heart disease > Visual impairment > Housebound > Peripheral vascular disease	9.6	1 (5.9)	Anaemia & haematinic deficiency > Weight loss & anorexia	0.7

In general, the comparison between similar and different sub-cohorts (gender and age groups respectively) is conducted at each frailty stages to answer research question three "*What are the significant patterns of frailty trajectories within cohort with point of concern within different elderly patient classification?"* The analysis between genders of all age sub-cohorts at frailty stage I (Fit > Mild) identified the combination of deficits polypharmacy and respiratory disease with both directions presents mostly in frailty stage I in all age sub-cohorts. The common pattern of frailty trajectories of Polypharmacy > Respiratory disease is the top two most followed in female elderly age sub-cohorts (65-74: n= 107, 15.2%), (75-84: n=89, 8.8%) and (85++: n=14, 5.9%) and exception of male in (85++: n=6, 6.5%). Meanwhile, male experienced pattern of frailty trajectories of opposite directional only in age sub-cohort (75-84: n=58, 9.0%). Furthermore, it is observed that, frailty trajectories of deficit diabetes is typically in frailty stage I with combination of deficits Polypharmacy or Visual impairment. The association of progression from Diabetes > Visual impairment is observed in early age of elderly sub-cohorts (65-74) or both male (n=65, 12.6%) and female (n=60, 8.5%) and (75-84) in male (n=62, 9.6%) only. The second association or frailty trajectories involving deficit Diabetes is Polypharmacy > Diabetes only in female of all age sub-cohorts; (65-74: n=110, 15.6%), (75-84: n=63, 6.2%) and (85++: n= 12, 5.0%).

Next, in frailty stage II (Mild > Moderate) it found that sub-cohorts (65-74) and (75-84) is usually associated with Dyspnoea and followed by either Heart failure or Respiratory disease. In male (65-74) both trajectories of Dyspnoea > Heart failure (n=17, 3.5%), Dyspnoea > Respiratory disease (n=15, 3.1%) and female (n=25, 6.0%) were found. Meanwhile, in sub-cohort (75-84) only male exhibits the frailty trajectories pattern of Dyspnoea > Heart failure (n=25, 4.1%). Female elderly from (65-74: n=28, 6.7%) and (75-84: n=31, 4.7%) sub-cohorts revealed trajectories of Osteoporosis followed by Polypharmacy. Only in sub-cohort (75-84) of both male (n=19, 3.1%) and female (n=41, 6.2%) exhibits the frailty trajectories of Visual impairment followed by Chronic kidney disease. However, in the last sub-cohort of (85++) all trajectories are unique.

In the third frailty stage, it revealed that frailty trajectories of Housebound > Requirement for care presents in sub-cohorts (65 – 74) male (n=11, 3.1%), female (n=16, 2.8%) and (75 – 84) female (n=28, 3.9%). The second apparent trajectories is Memory cognitive problem > Requirement for care in sub-cohort (65 – 74) male (n=10, 2.9%), female (n=13, 2.3%) and sub-cohort (75 –

84) male (n=13, 3.4%) and female (n=18, 2.5%). However, in sub-cohort (85++) the opposite directional of trajectories was identified only in male (n=1, 12.5%). In addition to that, only female in sub-cohort showed pattern of trajectories Housebound followed by Memory cognitive problem (65 – 74: n=15, 2.6%) and (75 – 84: n=21, 2.9%). Sub-cohort (75 – 85) showed trajectories of Fall followed by Fragility fracture in both male (n=9, 2.4%) and female (n=20, 2.8%).

Lastly, in the last frailty stage similar frailty trajectories as the previous stage was observed. Two trajectories (i) Housebound > Requirement for care of sub-cohort (75 – 84) both in male (n=3, 1.4%) and female (n=7, 1.6%) while the opposite direction appeared in female sub-cohort 85++ (n=4, 23.5%), while the opposite direction was observed in sub-cohort (65 – 74: n=6, 3.2%) in male. The second trajectories (ii) Memory cognitive problem > Requirement for care was observed in sub-cohort male (65 – 74: n=7, 3.8%) and female (85++: n=4, 23.5%) in opposite direction. In addition to that, trajectories Memory cognitive problem followed by Housebound was detected in sub-cohort (65 – 74: n=4, 2.2%) in male and the opposite direction in male (75 – 84: n=6, 2.8%). The final trajectories discovered Fragility fracture > Osteoporosis only in female of sub-cohort (65 – 74: n=6, 1.7%) and (75 – 84: n=7, 1.6%).

Research question four "*where do the shortest and longest transition of time occurred within cohort with point of concern*" was answered by examining the median duration of trace among the significant pattern of frailty trajectories. The frailty trajectories dominating the frailty stage I is Polypharmacy > Respiratory disease with the shortest duration observed in sub-cohort female (75 – 84) median duration of 3.6 months comprises in 89 traces (8.9%) while the longest median duration is in female sub-cohort of (65 – 74) with median 4.0 months consisting in 107 traces (15.2%). In frailty stage II, the shortest median duration was detected in female of sub-cohort (75 – 84) Osteoporosis > Polypharmacy with 0.7 month followed by 31 traces (4.7%), whereas the longest frailty trajectories (5.3 months) is in the same sub-cohort of male, Visual impairment followed by Chronic kidney disease with 19 traces (3.1%). Next frailty stage, showed that Fall > Fragility fracture is the shortest median duration (1.4 months) followed by 20 traces (2.8%) and the longest duration (28.2 months) is Requirement for care > Memory cognitive problem in sub-cohort (85++) male followed by 1 trace (12.5%). Finally, in frailty stage IV the (65 – 74) sub-cohort female with frailty trajectories Fragility fracture > Osteoporosis had the shortest median duration (0.6 month)

occurred in 6 traces (1.7%), while the Housebound had the longest median duration (16.7 months) before been diagnosed with deficit Requirement for care, recorded in 3 traces (1.4%).

6.3.5 Stage V: Evaluation

In this experiment, the evaluation stage was conducted mainly based on the analytical assessment along with the result from previous stage, IV mining and analysis. As the experiments aimed to investigate the progression of frailty based on the available clinical records, the calculation of rate of frailty progression and identification of different frailty definition to differentiate the rate of frailty progression which is highly beneficial. The approaches able to determine the point of concern in frailty progression which is high at different level of #frailty stages. One of the important findings gained in this experiment is `Polypharmacy` and `diabetes` are two deficits are commonly present in the early stage of frailty in male sub-cohort (65 – 75) with three significant frailty trajectories though female of same sub-cohort was recorded with two trajectories. While, the other sub-cohorts present with only one frailty trajectories associated with Diabetes with exception of sub-cohort (85++) in male.

The discussion with clinical experts suggested that rather than examining the whole thirty-six frailty deficits within the elderly, examining the deficit of concern among the frail elderly would be interesting. The deficits of concern are fall, hypertension and polypharmacy. The idea of focusing onto these frailty deficits are extremely useful as they are significant and prevalence within the frail elderly patient. Analysing these deficits and its association might give useful insights to domain experts in understanding the frailty progression.

6.4 Experiment 5: Association of Deficits of Concern with Frailty Progression

The analysis of frailty progression is explored in the previous section. It shown the pattern of progression on different frailty stages. The significant pattern of frailty trajectories involving high progression is valuable as it demonstrates the deficits that instigating a high rate of progression. It revealed that diabetes, polypharmacy, and hypertension are always resides within the early stage of frailty. Although fall and housebound are commonly found in the later stage, fall still provides an interesting frailty inquiry among the elderly with their frailty

progression. Hence, clinicians recommended a comprehensive and structural analysis of the three deficits.

The three frailty deficits; fall, hypertension and polypharmacy are the focus in this experiment as they are common within the elderly population (Cai and Calhoun, 2018; Jonas, Kazarski and Chernin, 2018; Liu *et al.*, 2020). Although, hypertension was identified as the highest prevalence among the three, its association with fall and polypharmacy is not unusual as reported in studies (Kojima *et al.*, 2011; Zaninotto *et al.*, 2020; Abu Bakar *et al.*, 2021). However, literature on the three deficits of concern and their association with frailty is very limited and only relying on the indicator of frailty at a single point of time (Bromfield *et al.*, 2017). Hence, implementing the widely acceptable frailty model as way to quantify frailty using longitudinal data to investigate the association between frailty progression with fall, hypertension and polypharmacy is inarguably intriguing.

This section will explore the association between deficits of concern. The association of the deficits will be further examined within the frailty stages and determine level of variability. It follows the general methodology explained in Chapter 3 and summarised from section 6.6.1 until 6.6.5.

6.4.1 Stage I: Planning

Planning stage is the start of the experiment where it involves the understanding the dataset for experiment and development of research questions. The aim of the experiment is to determine the association of deficits of concern within frailty stages. This experiment expands the previous study to answer three research questions:

1) *Can process mining detect and quantify the differences in frailty progression?*

2) *Is it possible to uncover the differences in sequence of deficits of concern using process mining?*

3) *Can process mining determine and evaluate the differences between patterns of concern?*

The hypothesis underlying these research questions is the discovered frailty progression model and its association with fall, hypertension and polypharmacy can be used as an initial step in analysing relatedness with frailty severity utilising process mining approach. The first research question addresses the temporality aspect of the frailty stages with respect to the presence of deficits of concern. It

focusses on the (RQ3), (RQ4) and (RQ5) of the research. Thee second research question investigates the fragmented pathway of concern. Lastly research question three explores the variation lies between patterns of concern. The last two research questions in this experiment address the (RQ1) and (RQ2) of the research.

6.4.2 Stage II: Extraction

Extraction was done using the SQL Server query database management system. Patient aged of at least 65 years old with their records of frailty deficits within study duration of 2003 – 2018 and timestamp details are selected. Apart from that, extra inclusion criteria are applied i) patient with at least one year of record within study duration, ii) patient aged over 85 years old whom their final frailty category is not in early stages of frailty and iii) maximum average accumulation of frailty deficits are three. The final exclusion criteria are excluded patient with a combination of any two of focused frailty deficits: fall, hypertension, and polypharmacy. The extraction stage has selected a total of 8,547 patients with three deficits of concern and 3,848 patients without any deficits of concern.

6.4.3 Stage III: Data Transformation and Loading

The transformation was done to create event log following the research questions specification. The first occurrence of the deficit only being considered in this experiment. The point of frailty stages within study duration consisting of fit (not frail), mild, moderate, and severe are identified based on the count of frailty deficits. Two additional steps were performed to create event log. The first is log enriching where the additional events are merged into the event log. Finally, the events with the same date are ordered alphabetically for frailty deficits while, for frailty stages, only the fit stage is arranged to be the first event following other deficits events and other frailty stages (i.e., mild, moderate, and severe) are located at the end of the deficit events. The rationale behind this approach is, fit stage is reached even when no frailty deficits recorded, but for other frailty stages, they required a minimum count of deficits to reach certain frailty stages. For mild, five deficits, moderate nine deficits and severe thirteen deficits.

The event log is loaded into two process mining tools in a structured way. First into the Disco tool to utilise its user-friendly interface and extracted event log in the format of .XES file. Next, the extracted event log is loaded into ProM tools for further analysis using process mining technique available in the form of plug-ins.

6.4.4 Stage IV: Mining and Analysis

Process mining with two perspectives and quantitative measures were used to conduct the mining and analysis in this study. Process discovery and conformance checking with two process mining perspectives; control flow case-based analysis and time perspective are included in this stage. Meanwhile, the quantitative measures were performed to analyse the frailty progression following the implementation of process cube and process variant analysis. Each of the steps will be discussed in detail below.

6.4.4.1 Analysis based Process Cube Analysis

A process cube is created from the generated event data. An example fragment of the processed event data set is shown in Table 6.8. A process cube is formed based on the events data with the following dimensions: frailty stages, year (in one-year increment), gender, age, and clinical concept. It is possible for the process cube to have several number of dimensions to disseminate process models and event logs across numerous cells in process cube (van der Aalst, 2013). The large event data is being characterised by the process cube structure which is linked to patients' properties and events. Next, a process cube view is constructed based on a process cube structure which defines which dimensions and events are selected for analysis. It can be done using the process cube operators such as slicing, dicing, drilling down, and rolling up.

Table 6. 8 An example fragment of event data of one patient after processing steps

ID	ACTIVITY	FRAILTY STAGES	CLINICAL CONCEPT	TIMESTAMP
1	Fit	Fit	No	25/07/2001 21:08
1	Arthritis	Fit[1]	Disease State	25/02/2014 20:39
1	Activity Limitation	Fit[2]	Disability	15/09/2000 02:42
1	Anaemia	Fit[3]	Abnormal Lab Value	14/07/2011 05:06
1	Atrial Fibrillation	Fit[4]	Disease State	13/11/2002 17:54
1	Mild	Mild	No	29/11/2006 21:00
1	Cerebrovascular Disease	Mild[5]	Disease State	23/03/2001 21:58
1	Chronic Kidney Disease	Mild[6]	Disease State	25/05/2006 03:07
1	Diabetes	Mild[7]	Disease State	10/08/2001 06:45
1	Dizziness	Mild[8]	Symptom/Sign	14/05/2010 16:14
1	Moderate	Moderate	No	24/11/2010 01:53
1	Dyspnoea	Moderate[9]	Symptom/Sign	31/10/2014 02:14
1	Fall	Moderate[10]	Symptom/Sign	25/07/2002 12:01
1	Foot Problem	Moderate[11]	Disease State	08/10/2013 12:48
1	Fragility Fracture	Moderate[12]	Disease State	25/05/2012 02:58
1	Severe	Severe	No	12/08/2013 09:57
1	Hearing Impairment	Severe[13]	Disability	15/08/2003 08:30
1	.	.	.	
1	.	.	.	
1	.	.	.	

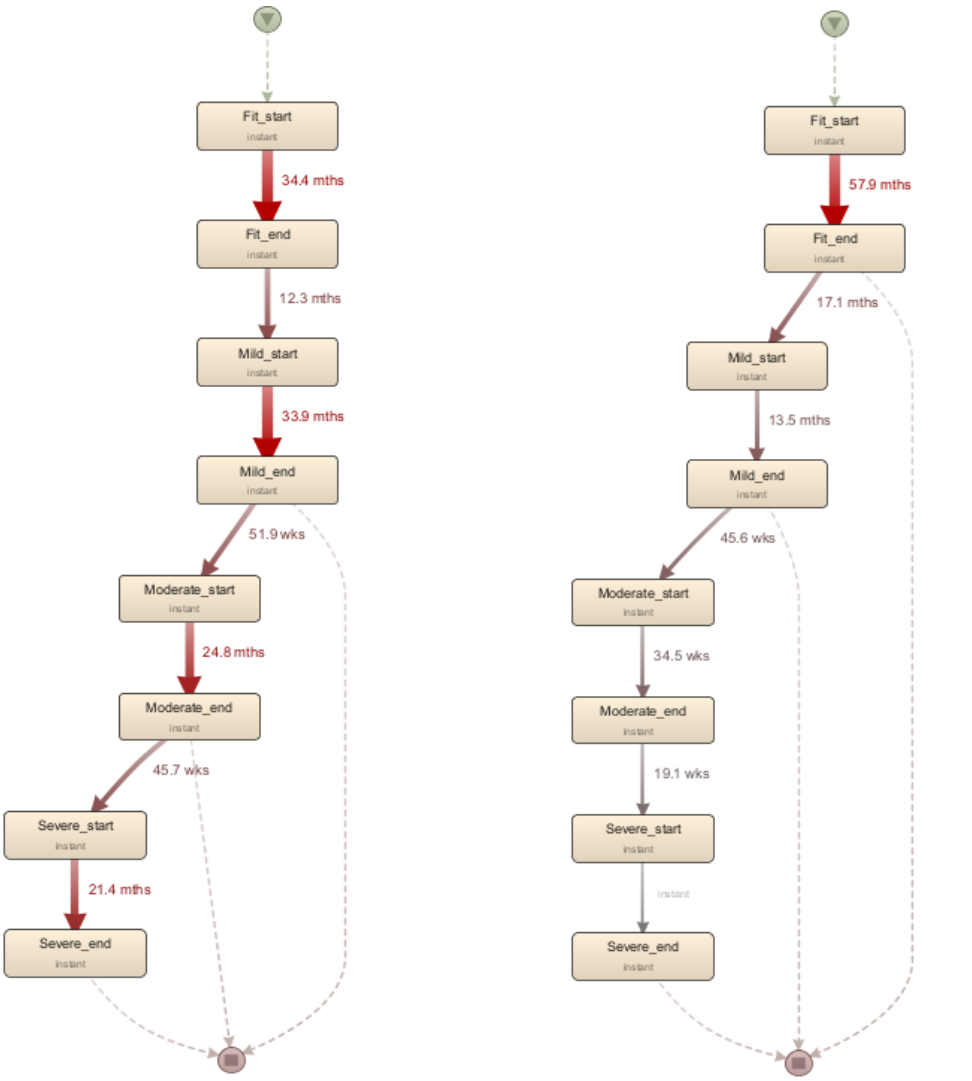
1	Weight loss anorexia	Severe[36]	Symptom/Sign	15/08/2003 08:30

The dice process cube operator is applied to a frailty stage dimension. Using dicing operator, a set of values; Fit, Mild, Moderate and Severe will only be considered. Note that, dicing operation is only restricts the sets of values resides within dimensions and not removing it. This results in process view comprises of sets value of; fit, mild, moderate, and severe in dimension frailty stages. Two event logs are formed, and the descriptive statistics is shown in Table 6.9. Process conformance checking with quality dimensions of trace fitness, precision and generalisation are measured on both event logs. Both models are highly representative from the event log where Figure 6.2 (a) replay fitness is 0.89, a precision of 0.99 and generalization of 0.833, while Figure 6.2 (b) replay fitness is 0.87, precision of 0.97 and generalization of 0.53. However, only generalisation measurement on sub-cohort without deficits of concern is medium (0.53) indicates that the model is in medium level of estimating the future behaviour of the process.

Table 6. 9 The descriptive statistics of sub-cohorts with and without the deficits of concern

Name of Interval	Cohort with Deficits of Concern	Cohort without Deficits of Concern
# Patient	8,547	3,848
# Events	30,754	5,385
Events per patient	3.6 (~4)	1.4 (~1)
Fitness	0.89	0.87
Precision	0.99	0.97
Generalisation	0.83	0.53

Process discovery of the two sub-cohorts is illustrated in Figure 6.2 using the performance view generated from Disco. The edges show the average duration within frailty stages or the transition point. Transition point defined as the point of starting frailty stages until the end of the frailty stage, example for fit stage, the average duration it took between Fit_start to Fit_end is 34.4 months from Figure 6.2(a). While for the transition point between frailty stage fit to mild, represents in Figure 6.2(a) as Fit_end to Mild_start took about 12.3 months. The sub-cohort help to reveal the relationship between the status of deficits of concern and how it linked with frailty progression.



a) With deficits of concern

b) Without deficits of concern

Figure 6. 2 The process model of the two sub-cohort showing the flow of frailty progression. The boxes represent the stages of frailty with the edges indicates the average duration took since the end of previous activity to the start of the next activity.

Figure 6.2 shows the flow of frailty stages using performance view following the mean duration of process metric. The detail of both process models is showing 100% activities with 100% frequent paths. The activity of the model is referring to the frailty stages identified based on the deficit count. The relationship of the two sub-cohorts were inspected at each frailty stages. The hypothetical testing was applied using independent t-test at each stage with the general hypothesis is time taken to reach the subsequent frailty stages is influenced by the deficits of concern. A null hypothesis state that time taken to reach the next frailty stage in patients with the combination of three deficits of concern are longer than patient without the three deficits of concern.

Table 6. 10 Descriptive statistics of two sub-cohort (with deficit of concern, n = 8,547 and without deficits of concern, n = 3,848) within frailty stages represents the duration within each stage (in months)

Frailty Stages / Transition Point (TP)	n	Case Duration, Mean (Min - Max) [Months]	Case Duration, Median (IQR) [Months]	n	Case Duration, Mean (Min- Max) [Months]	Case Duration, Median (IQR) [Months]	P-Value
	With Deficits of Concern			Without Deficits of Concern			
Fit	8,547	34.4 (0 – 186.0)	26.6 (405 – 1,434)	**3,848**	**57.9 (0.1 – 180.5)**	**46.6 (25.0 – 82.0)**	**0.00**
Mild	**8,514**	**33.9 (0 – 166.7)**	**27.8 (14.3 – 47.8)**	1,432	13.5 (0 – 4,586)	1.0 (0 – 18.6)	**0.00**
Moderate	**7,023**	**24.8 (0 – 146.2)**	**19.2 (5.8 – 36.6)**	101	7.9 (0 – 91.8)	0 (0 – 9.9)	**0.00**
Severe	3,335	21.4 (0 – 143.4)	11.1 (0 – 33.7)	2	0 (0 – 0)	0 (0 – 0)	0.25
TP I	8,514	12.3 (0 – 149.6)	7.0 (2.1 – 16.6)	**1,432**	**17.1 (0 – 129.8)**	**9.0 (2.6 – 23.6)**	**0.00**
TP 2	7,023	12.0 (0 – 104.6)	6.7 (2.0 – 16.5)	101	10.5 (0 – 88.8)	5.5 (1.7 – 14.6)	0.32
TP 3	3,335	10.5 (0 – 92.6)	6.0 (1.9 – 14.6)	2	4.4 (2.8 – 6.0)	4.4 (3.6 – 5.2)	0.49

Table 6.10 demonstrates the numerical information generated from the statistical comparison between sub-cohorts with and without the three deficits of concern. The highlighted value in Table 6.11 represents the higher duration between the sub-cohort with *p*-value of less than 0.05.

In general, the statistical significance difference showing difference in frailty stages fit, mild and moderate, while significance difference only found in transition point 1 between the two sub-cohorts. The statistical comparison is showing significance difference when *p*-values is less than the threshold value 0.05. It is observed that the average duration in the sub-cohort without the three deficits of concern is lower than the cohort with three deficits of concern start from stage Mild frailty stages and onwards. It is expected as the proportion of patient between the sub-cohorts is differ by 35% and the percentage differences of patient number at each frailty stages; fit stage (35%), mild stage (70%), moderate stage (93%) and severe stage (98%). The percentage differences value indicates that the differences between the two sub-cohort is higher towards the end of frailty stage.

Consequently, the model in Figure 6.3 presents the lower average duration between the frailty stages of sub-cohort without the deficits of concern. The mean and median case duration was also reflected to the values as the cohort with deficits of concern is 10.3 years (10.7 years) and sub-cohort without deficits of concern is 5.8 years (4.8 years). However, only in fit stage the sub-cohort with

three deficits of concern spent less time than the other sub-cohort as well as the time taken in transition point 1,

While there is little doubt that treating frail elderly with hypertension has a cardiovascular benefit, evidence from studies have prompted concerns that the treatment in elderly may potentially have side consequences such as dangerous falls (Callisaya *et al.*, 2014; Tinetti *et al.*, 2014). In the next section of this experiment, the sequence between risk of fall with hypertension and polypharmacy will be explored. The association will be analysed by comparing sequence between them and frailty progression. In the dimension selection step, only deficits of concern were selected as first exploration step for deficits of concern sequence analysis.

Two process cube operations were applied to the original process cube structure in exploring the sequence within the deficits of concern. The first operation is slicing operation using clinical concept dimension and resulted in events that relate to disease state are considered. After that, dicing operation on activity dimension is applied where falls, hypertension and polypharmacy are considered as deficits of concern to be in the process cube view. An event log will be created using the process view

Figure 6.3 shows the events sequence for three deficits of concern using the trace variant. It shows that the trace variants following three unique patterns based on the first event, fall occurred to an elderly person. The first sequence pattern where fall occurred in later sequence after hypertension and polypharmacy, the second sequence pattern is where fall occurred in between either hypertension or polypharmacy, the last sequence pattern is where fall occurred before hypertension and polypharmacy. The first sequenced pattern dominated the sub-cohort with 5,256 case (61.5%), following the second with 2,244 cases (22.5%) and the third with 1,047 cases (16.0%).

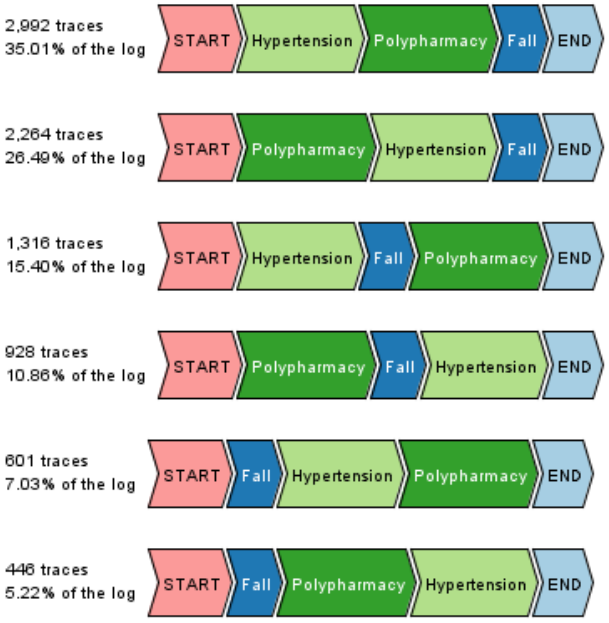


Figure 6. 3 Trace variant analysis generated using ProM

An Analysis of Variance (ANOVA) was performed on all parts to test for a difference of sequence pattern on case duration. Threshold p-value of 0.05 is set as frequently reported in literature to assume that the value of outcome from tested parts is statistically significance. Later, post-hoc test, Tukey significance difference was conducted to find which part is different. The significance tests were used with hypothesis that there are differences among sequence pattern with its case duration. Table 6.11 shows the descriptive statistics associated with sequence pattern.

Table 6. 11 The numerical information generated by statistical tests

Pattern of sequence	Trace Variant	# Patients (%)	# Events	Case Duration, Mean (Min- Max) [Days]	Case Duration, Median (IQR) [Days]
I	1 & 2	5,256 (61.5)	32,927	3,858 (378 – 5,684)	4,065 (2,964 – 4,947)
II	3 & 4	2,244 (26.3)	13,749	3,598 (382 – 5,752)	3,729 (2,623 – 4,656)
III	5 & 6	1,047 (12.2)	6,384	3,478 (429 – 5,756)	3,660 (2,375 – 4,558)

The average case duration of 3,858 days (10.5 years) at 95% CI (3,824 – 3,892) days or (10.4, 10.6) years for sequence pattern I; 3,598 days (9.9 years) at 95% CI (3,546 – 3,651) days or (9.7, 10.0) years for sequence pattern II; and 3,478 days (9.5 years) at 95% CI (3,398 – 3,557) days or (9.3, 9.8) years for sequence pattern III are observed. There is statistical significance difference between the sequence patterns and their case duration with P-value is 0.00. Post hoc comparisons using Tukey test were carried out. The result shows that the case duration for all three patterns is statistically significance when considering its

pattern based on sequence of deficit of concern. Furthermore, sequence pattern for fall occurred after both hypertension and polypharmacy recorded the longest case duration with median of 4,065 days or 11.1 years compared to the other sequence pattern.

The patient following the pattern of sequence I, II and III is identified to create three separate event logs. It reveals that based on 50% frequent path that combination or an individual of deficits hypertension and polypharmacy occurred before reaching mild frailty stage in patterns of sequence I and II which affecting about 84% of cases. Apart from that, 50% traces of first fall event happened before reaching mild frailty stage in pattern of sequence II and III. While this situation is opposite in pattern of sequence I. Furthermore, the average duration of reaching severe stage is the shortest in pattern of sequence I (compared to pattern of sequence II and III) which took about 14 months. This transition observed after fall had occurred.

6.4.4.2 Variant Analysis

The dominant sequence pattern comprises of two top variants of the same process (frailty progression) or pattern of sequence I is further analysed. The variants will be compared with each other to identify different behaviour recorded in the event logs. The variant analysis was applied in this study using the plug-in implemented in the ProM called Process Comparator (Bolt, De Leoni and Van Der Aalst, 2016; Bolt, de Leoni and van der Aalst, 2018). The two event logs will be compared to produce comparison results in a form of transitions model as discussed in Chapter 3 in Section 3.2.4.1.2.

The second stage dimension selection was done as part of the event log enriching step. It used to create the event log from the two variants comprises of the three deficits of concern and frailty stages. Table 6.12 shows the descriptive information of the two event logs for variant analysis based on pattern of concern.

Table 6. 12 The descriptive information of event logs from the top two variants

Characteristics	Variant I	Variant II
# Patient (%)	2,564 (58.5)	1,821 (41.5)
# Events (%)	15,817 (57.7)	11,581 (42.3)
# Trace Variant	29	33
# Activities	7	7
Activities per Patient, Mean [Min -Max]	6 [4 – 7]	6 [4 – 7]
Case Duration, Mean [Min- Max] (days)	3,922 [432 – 5,682]	3,821 [378 – 5,684]
Case Duration, Median [IQR] (Days)	4,134 [3,074 – 4,949]	3,999 [2,893 – 4,954]

Pattern of concern is defined as a sequence of deficits concern that generate a unique pattern. Table 6.12 shows similar percentage dominated by variant I, 60% of patient and event out of sequence pattern I. The number of variants is larger

in variant II (33) suggesting that variability presents between the two variants. The pair-wise difference percentage produced by the Process Comparator plug-in is 68.97%.

Two abstraction types are available with the Process Comparator Plug-in to create the transition system. The first type where it considers only the last event of traces and second is the last two events. The first type of abstraction produced a directly followed model which other process mining tools such as Disco created. However, often the complexity is high due to loop formation which may limit the mining accuracy of class of processes (Van der Aalst and Weijters, 2004).

Figure 6.4 shows the joint control-flow of two variants using trace frequency for process metrics. The second type of abstraction was implemented in the joint control-flow to unfold loops (which commonly appear using first type of abstraction) into sequential flow. The state of model represents the sequence of last two activities of traces where first name of state is the last activity in the sequence, e.g., in state *"[Hyper],[Mild]*" last activity is *[Hyper]* and *[Mild]* is next in the sequence.

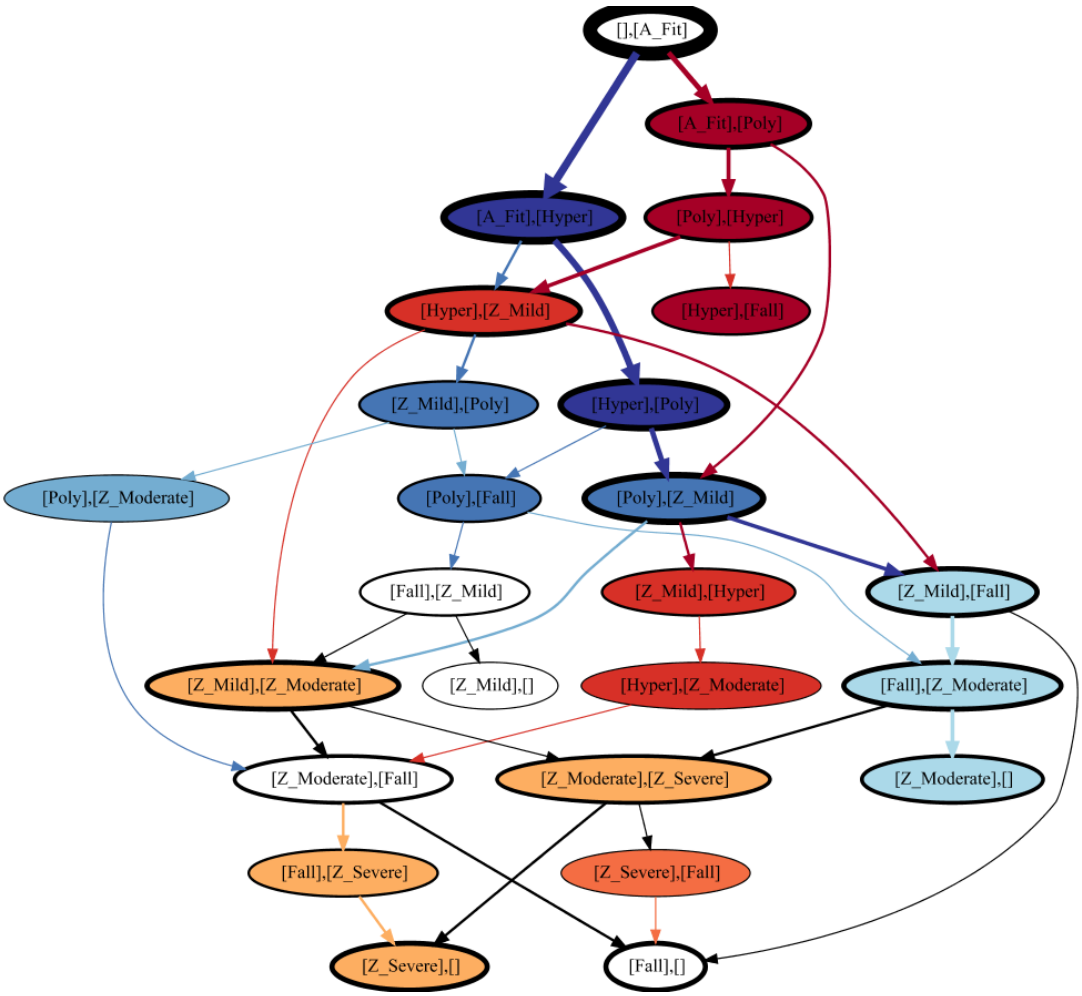


Figure 6. 4 Transition system of two variants using the second type of abstraction with process metrics: trace frequency

There are behaviours which its frequency occurrence is high statistically significant only in single dominant state. The sequence states of *"[Fit],[Hypertension]*" and "*[Hypertension],[Polypharmacy]*" occurred at 97.39% and 72.5% only at variant 1. It indicates that these states are only prevalence in Variant 1. On the other hand, the single dominant states for Variant 2 are *"[Fit],[Polypharmacy]*", "*[Polypharmacy],[Hypertension]*" and "*[Hypertension],[Fall]*" occurred at 95.5%, 68.48% and 24.77%. The similar pattern can be observed at this point is, both deficits of concern; Hypertension and Polypharmacy frequently happened before mild frailty stage. The frequency average for Variant 1 is 56.71% while in Variant 2 is 53.32%.

Next observations can be derived on the third deficit of concern, fall. The frequency occurrence involving fall in Variant 1 (blue-based colour states) is significantly higher in the following path: "*[Polypharmacy],[Mild]*" (56.71% vs 29.87%), "*[Mild],[Fall]*" (38.30% vs 31.63%), "*[Fall],[Moderate]*" (38.77% vs 31.85%) and "*[Moderate],[]*" (32.18% vs 25.37%). Meanwhile in Variant 2, three

states which present high statistically significant difference in frequency occurrence in *"[Mild],[Moderate]*" (34.71% vs 26.72%), "*[Fall],[Severe]*" (23.39% vs 15.95%) and "*[Severe],[]*" (37.29% vs 29.41%).

The comparison of variant using the Process Comparator Plugin reveals a difference between variants which is pattern for deficit of concern fall. In Variant 1, fall is statistically significant occurred in between Mild and Moderate frailty stages. Whereas, in Variant 2 high statistically significant path for fall happened before severe stage.

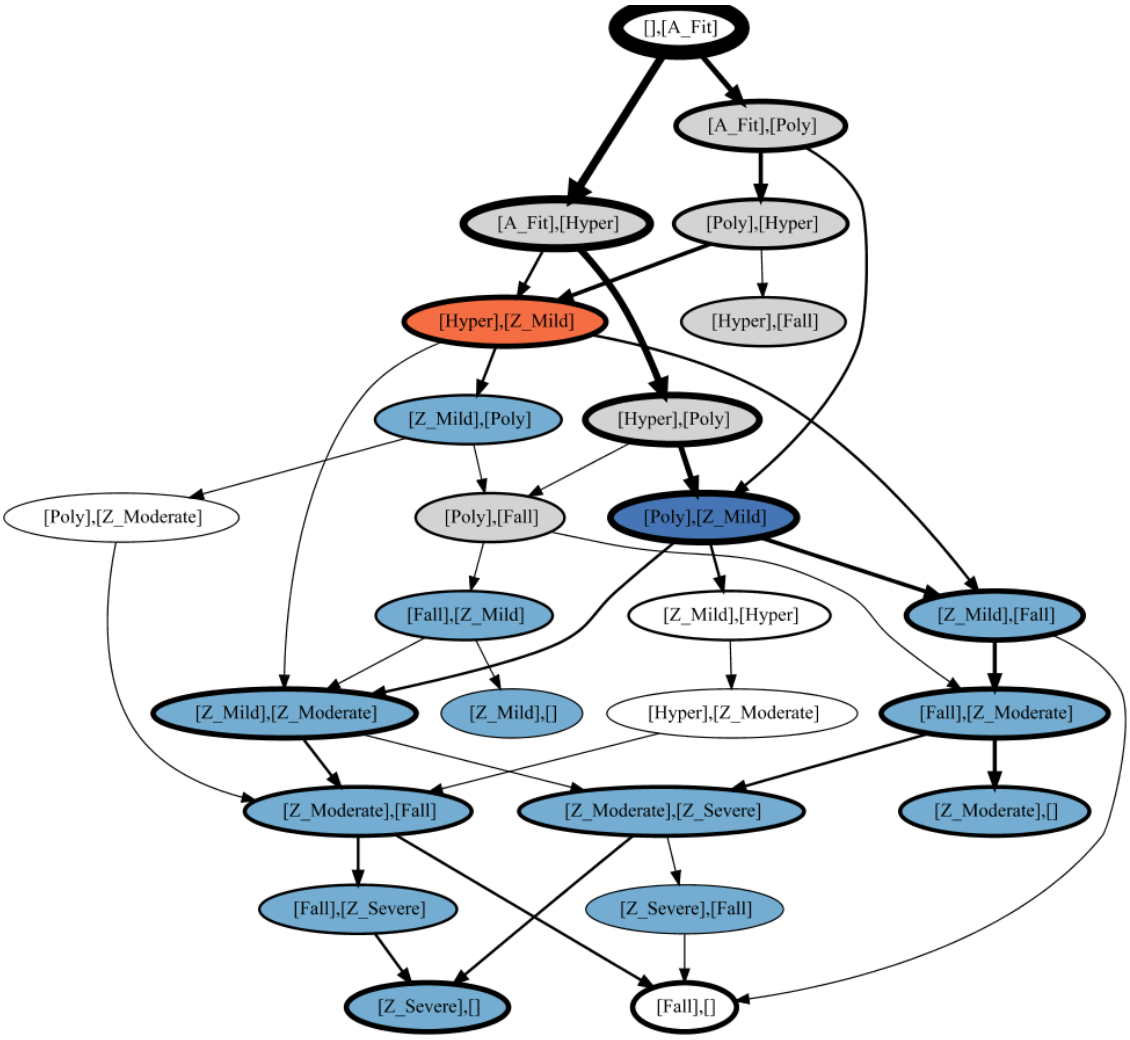


Figure 6. 5 Transition system of two variants using the second type of abstraction with process metrics: elapsed time

In addition to trace frequency for process metrics, comparison of elapsed time between the two variants was conducted. Figure 6.5 illustrates the transition system following the comparison setting of elapsed time. The result shows that generally, more states label is statistically significant differences is high in Variant

1. However, only one state is statistically significant higher in Variant 2 “[Hypertension],[Mild]” (723 days vs 444 days).

Table 6.12 shows the statistically significance of measured elapsed time of state label. The highlighted average elapsed time represent the statistically significance difference state label for certain variant. The statistically significant state highlighted the longer elapsed time to capture the delay between two activities. However, in the frailty progression concept, the delay in transition to other activities are consider as good progression. Thus, from the numerical information provided in Table 6.13 shows that Variant 2 elapsed time is statistically lower than Variant 1 in all state label except for “*[Hypertension],[Mild]*”.

Table 6. 13 Comparison of average elapsed time between the two variants.

State Label	Variant I	Variant II
	Average Elapsed time (days)	
“[Polypharmacy], [Mild]”	**992**	415
“[Hypertension], [Mild]”	444	**723**
“[Mild], [Fall]”	**1,695**	1,420
“[Fall], [Moderate]”	**2,738**	2,339
“[Mild], [Moderate]”	**1,335**	1,101
“[Moderate], []”	**3,395**	3,083
“[Moderate], [Fall]”	**2,803**	2,421
“[Moderate], [Severe]”	**2,403**	2,009
“[Severe], []”	**3,777**	3,506

6.4.5 Stage V: Evaluation

The feasibility of the approach chosen in analysing the association of deficits of concern with frailty progression was examined as part of the evaluation. The step conducted by diagnosing the findings of analysis and relate to the relevance of approaches taken. The statistical and analytical assessments were presented as part of the findings in mining and analysis section.

Comparison analysis of deficits of concern with frailty progression is the aim of the experiment. The formation of research questions incorporating process cube approach has established a basis in examining the rest of questions starting with general question (in RQ1). Apart from that, the approach managed to demonstrate the ability of the approach in analysing complex clinical data using combination of levels of granularity of events referring to the case attributes throughout the experiment.

The finding was strengthened by integrating the concept of statistical significance analysis as it crucial in gaining clinician trust in analysing using clinical data. The interpretation of significance differences was established using the widely used

statistical measurement of *P*-value. Although, other element of comparison can be made to discover the differences between cohorts of patients, the case duration and duration between stages was ideal in analysing progression.

From clinical part, the differences between sub-cohorts of patients with specified attributes were formally reported. In general, it is observed that the case duration in patient with deficits of concern is five years longer (for median case duration) compared to another sub-cohort. While pattern of sequence I had the longest case duration but with the shortest interval to reach severe stage, it suggests that patient following the pattern are becoming severe more quickly. It is mainly happened after fall had reached moderate frailty stage first.

Apart from that, frail elderly patients who have been prescribed with more than five medications are highly likely to be associated with other morbidities such as chronic kidney disease, heart disease, diabetes (Aubert *et al.*, 2016). The condition is increasing the number of deficits accumulated which demands substantial commitments for diagnostics investigation in outpatient care and other use of services in hospital (Buja *et al.*, 2020). This reflects the finding that severe frailty stage patients are at high risk of longer hospital stays and multiple hospital admission (Clegg *et al.*, 2016). Consequentially, this makes the case duration high due to clinical commitments.

The experiment provides proof on the possibility of investigating frailty progression with each stage in relation to deficits of concern. The variation of pattern was easily determined by the application of the process cube approach in process mining. Apart from that, external feedbacks from the process mining committee suggest looking into the possibility of utilising the occurrence of recorded data for exploring the level of deficits severity.

The discussion with clinical experts were concluded that the visualizations reveal the significant difference of sequences between deficits of concern within different frailty categories. The approach in integrating process mining in examining the interaction of deficits of concern with frailty progression is interesting and valuable for the analysis of frailty. The extension of this approach was suggested as potential future work is to include other deficits prevalence in frail elderly such. The other deficits of concerns could be mapped to create as the recommended visualization and as guidance in the frailty study.

6.5 Summary

Three experiments have been presented in this chapter using the Connected Bradford dataset. The first experiment explored the significant trajectories employing the statistically significant approach within different frailty category of frail elderly. The second experiment investigated the different rate of frailty progression and its significant pattern within frailty stages. The last experiment analysed the association of deficit of concern with frailty progression.

According to the clinicians, process models created using process mining techniques are valuable in visualising the trajectories in understanding frailty within elderly. For example, in experiment 3 and 4 in this case study revealed the trajectories representation in several visualisations including Directly Followed visual Miner in Figure 6.1, trace variant in Figure 6.2 and trajectories patterns of directly followed in Figure 6.8. Meanwhile, based on discussion with clinician, focus analysis on the deficits of concern in association with frailty progression in Experiment 5 is potentially useful in demonstrating which deficits of concern has high influence on how the frailty progress. This has been presented in Figure 6.5.

The works (experiment #3) in this chapter is an addition in answering research question two of thesis, "*is it possible to apply process mining to model the frailty trajectories?"* The main work focused in answering research question three, "*what is the progression of frailty?*" in experiment #4 and research question four "*what is the association between deficit of concern and frailty progression?*" in experiment #5.

The next chapter of the thesis will explore the confirmatory analysis of the cut off point for frailty categories. The chapter serves as the comparison analysis in employing different approach in determining the interval for frailty stages.

Chapter 7

Confirmatory Analysis: Electronic Frailty Index Score

Analysis of the Connected Bradford dataset provides various insightful results which have been presented in Chapter 6. This chapter presents a confirmatory analysis on the application of the eFI score within the General Practitioners (GP) primary care system SystmOne. The analysis explores machine learning approach in determining the cut off points of accumulated deficits. The initial part of the chapter involves exploratory data analysis of the Connected Bradford dataset. The later sections investigate the comparison of categories of elderly between machine learning approach and the adopted eFI from literature (Clegg's approach).

7.1 Motivation of Analysis

The electronic Frailty Index (eFI) score is a tool to measure the severity of frailty in elderly people (Clegg *et al.*, 2016). It has been adopted throughout the primary care system, SystmOne to automatically calculated the score using routinely collected health record. Frailty index score initially intended to be used as a continuous score, however the "stratum-specific likelihood ratio" or categorisation of score has been established for comparison purpose with frailty phenotype (Hubbard, O'Mahony and Woodhouse, 2008). The eFI score stratify frail elderly into four categories of fit or not frail, mild, moderate, and severe.

The categorisation of frailty categories was based on the cut-off points or score. Several studies have reported various cut off points to differentiate between frail and non-frail elderly. A cut-off point of 0.2 was established in (Searle *et al.*, 2008), while a cut-off point 0.25 was most commonly adopted (Rockwood, Andrew and Mitnitski, 2007; Song, Mitnitski and Rockwood, 2010; Eeles *et al.*, 2012; Wou *et al.*, 2013). The eFI score was defined based on the quantile of 99^{th} centile as the upper limit following the 0.25 cut-off points (Clegg *et al.*, 2016). The cut-off points of 0.25 was based on the crossing point in density distribution of deficits between robust group and frail group (Rockwood, Andrew and Mitnitski, 2007). Nevertheless, it is solely based on the accumulated deficits in population distribution at the end of study duration with no other variables considered.

A large and growing body of literature following the eFI approach within the elderly research has been done (Lansbury *et al.*, 2017; Ambagtsheer *et al.*, 2019; Fogg *et al.*, 2022). In addition to that, the application of the eFI in the healthcare system and research work has been validated by numerous studies focusing on the effectiveness of eFI implementation in segregating elderly for better healthcare interventions (Brundle *et al.*, 2018; Stow *et al.*, 2018; Abbasi *et al.*, 2019; Broad *et al.*, 2020; Callahan *et al.*, 2021).

Although, eFI is widely recognized to be sensitive in differentiating several degrees of frailty among the elderly, the cut-offs of frailty score are still debateable. The continuous frailty index has predictive accuracy of moderate to good while the dichotomized frailty index has slightly lower predictive accuracy (Hoogendijk *et al.*, 2020) as a result that the frailty index was created initially in continuous form without the cutting-off points (Rockwood, Andrew and Mitnitski, 2007). A few studies has addressed this issues by proposing the new cut off points of frailty index by taking into account the age (Romero-Ortuno, 2013) and investigating novel cut-off points based on gender specific (Hoogendijk *et al.*, 2020). However, other frailty measurement properties need to be included to determine the optimum cut-off points of frailty index.

7.2 Exploratory Data Analysis of Bradford Dataset

The inspection of the log is performed on the processed event log. It involves the Exploratory Data Analysis (EDA) method introduced in the classic work of (Tukey, 1997) as building initial knowledge or insight, interpreting the dataset, creating a critical hypothesis to support domain study theory using the dataset employing a graphical approach. The graphical approach usually acts as the primary component in the EDA (Morgenthaler, 2009) as a guide to look at the dataset from various points of view. The EDA exhibits an essential element in revealing the data's fundamental phenomena, followed by the confirmatory analysis recognised as the inductive approach (Jebb, Parrigon and Woo, 2017).

The purpose of visualizations generated at this stage is to show the distribution of cohorts in the study population. They usually require the integration between multiple tables or calculated values of entries in categories within tables. The procedures of generating the visualizations as the log profile figure begins as the complex queries executed within the SQL server before transformation steps take place in the Python Jupyter Notebooks. Such a systematic approach is taken as the author has limited authority in the SQL environment to only writing SQL but

not to create additional tables or the outcome tables and to ensure we are working on the same cohort throughout the research.

The following section describes the study population's initial examination from Section 4.8.1, Table 4.14, based on the EDA method adopted from (Bezerra *et al.*, 2019) using visualization tools such as Tableau and Jupyter Notebooks. It involves the visual analysis of data and interaction between categories within the cohort.

7.2.1 Visual Analysis of Single Attributes/Variables

The dataset is composed of a combination of categorical or qualitative and numerical or quantitative variables. Quantitative assessment and variable proportion are several of the approach of EDA for presenting the categorical occurrences across different category.

7.2.1.1 The Cohort Distribution

The patient included in this study was following the selection criteria in Section 5.9. It illustrates the distribution of the patient and both genders at the start of the study. The cohort distribution in Figure 7.1 represented by the pyramid population illustrates the structure of the cohort by comparing the relative number of patients within different age groups. The age of the patient was grouped into five years range based on the first record appear within the study duration. There are 72 patients (0.008%) are eliminated because their gender is not represented in the figure. There were 17.14% more female patients than male in the cohort with 50,671 female patients (58.7%) and 35,844 male patients (41.4%).

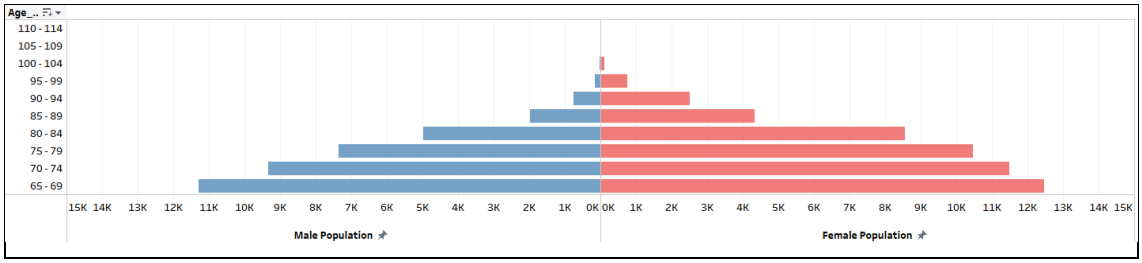


Figure 7. 1 The pyramid population of the Bradford dataset. The y-axis shows the five-years age range and the x-axis represents the number of patients in thousand. The female patient is coloured coded with light red in the right side whereas for male patient is on the left side

The frailty category of the cohort was determined using the electronic frailty index score in Section 5.7.2. The final score within the study duration was presented in the Figure 7.2 with the five-years range and both genders at the end of the study.

The figure presents the distribution of the cohort within the frailty category where most of the cohort appear to move into the (80-84) and (90-94) year range groups.

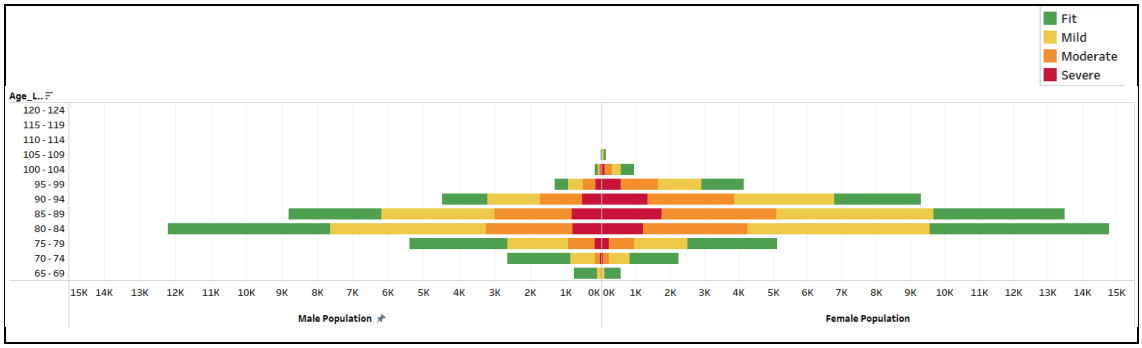


Figure 7. 2 The pyramid population of the Bradford dataset. The y-axis shows the five-years age range from the 65 until 124 and the x-axis represents the number of patients in thousand. The colour represents the frailty category with the male patient positioned at left side of figure.

7.2.1.2 Frailty Deficits Distribution

The maximum number of deficits associated with frailty is 36 and the Figure 7.3 is presenting the distribution of number of deficits acquired within each frailty category within the study duration. While it is known from the past work that the clinically fit category should constitute the highest number of patients in the cohort, but the visualization is very useful in describing in detail the proportion of patient in each deficit count and the maximum number of deficits the cohort recorded. Each frailty category constitutes about 36.8% for clinically fit, 33.09% for mild, 21.1% for moderate and only 9.0% for the severe. The visualisation encapsulates that highest number of patients in the cohort 7,571 (8.75%) resides in the clinically fit category with no frailty deficit recorded. The trend of the figure shows that the number of patients is increasing with the count of unique deficit starting from one until five and it gradually decreasing after that.

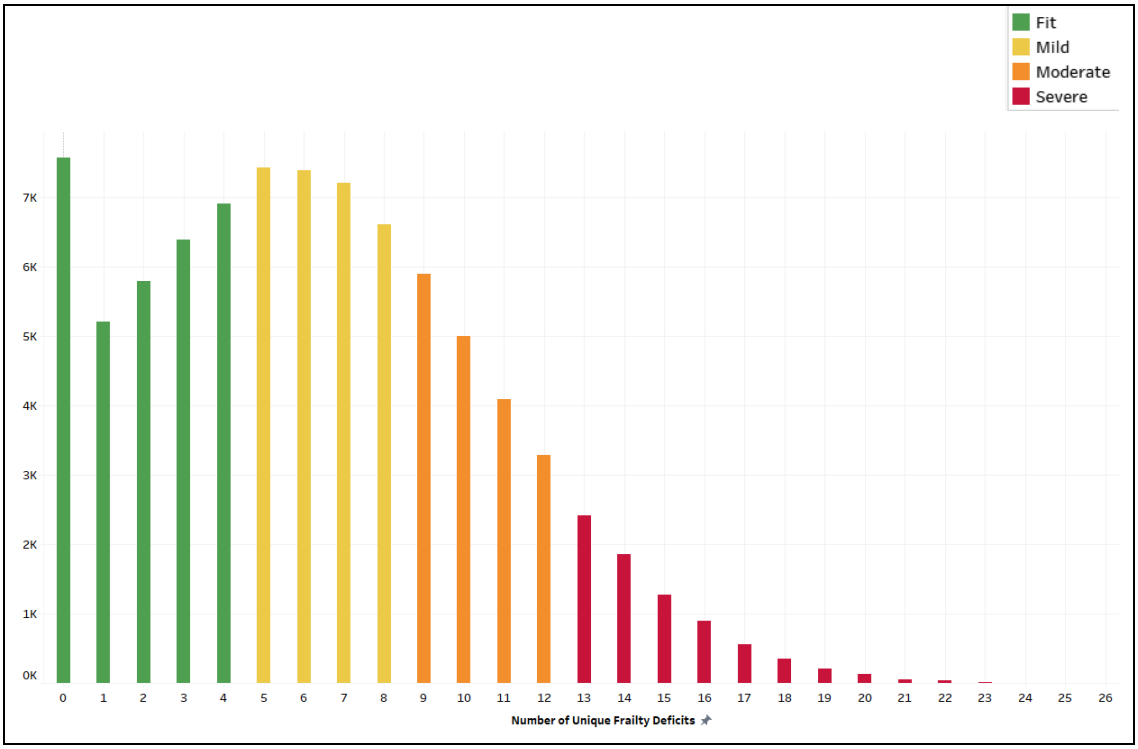


Figure 7. 3 The distribution of the number of unique deficits. The y-axis represents the number of patients while x-axis represents the count of unique frailty deficits.

The next visualisation is a treemaps shown in Figure 7.4. It demonstrates the value of unique frailty deficit using the hierarchical illustration within the study cohort. Figure 7.4 shows the proportion of which frailty deficits that mostly associated with elderly. Poly indicates Polypharmacy, Hyper is Hypertension, VI is Visual Impairment, House is Housebound, USD is Urinary System Disease, A is Arthritis, CKD is Chronic Kidney Disease, RD is Respiratory Disease, Fall is any event associated with elderly falling, Dysp is Dyspnoea, MCP is Memory Cognitive Problem, IHD Ischaemic Heart Disease, Diab is Diabetes, AHD is Anaemia Haematinic Deficiency, AF is Atrial Fibrillation, HI is Hearing Impairment, CD is Cardiovascular Disease, Hypo is Hypotension, Dizzy is Dizziness, Osteo is Osteoporosis, FF is Fragility Fracture, UI is Urinary Incontinence, RFC is Requirement for Care, MTP is Mobility Transport Problem, HF is Heart Failure, SV is Social Vulnerability, SU is Skin Ulcer, TD is Thyroid Disease, WLA is Weight Loss and Anorexia, PVD is Peripheral Vascular Disease, SD is Sleep Disturbance and FP is Foot Problem. The size of the rectangles represents the relative size of data category (frailty deficits) with the count of unique deficit and its percentage. The polypharmacy deficit contributes to the largest proportion (starts from the top left corner) within the cohort following with the hypertension, visual impairment, housebound and urinary system disease.

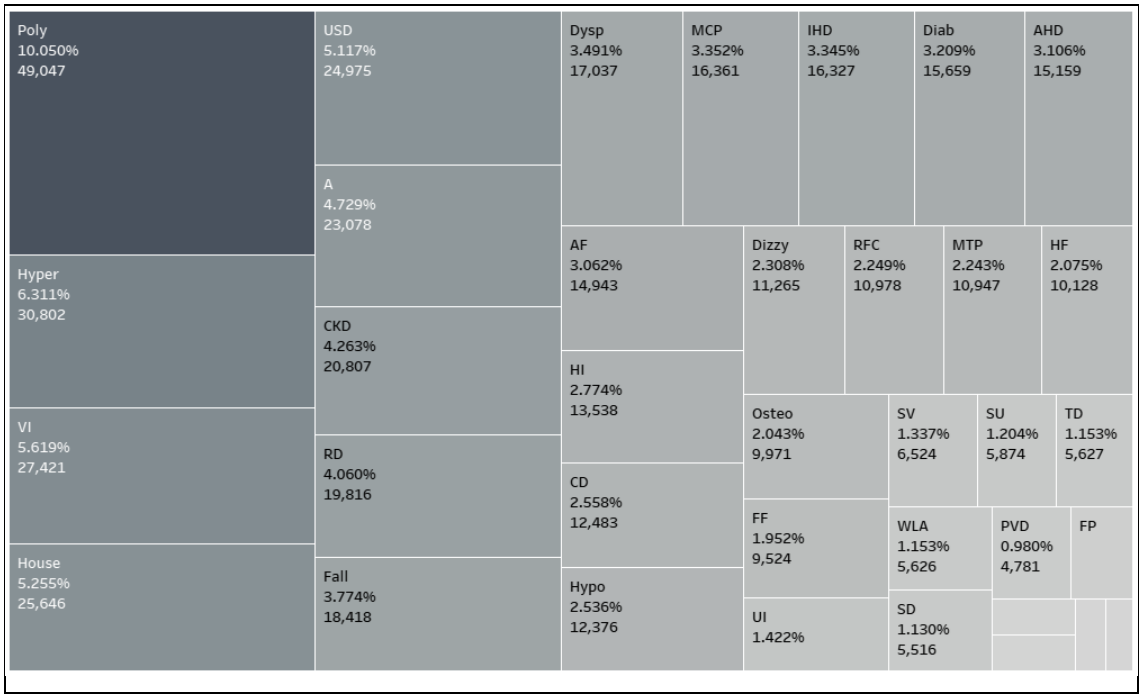


Figure 7. 4 The treemaps encapsulates the proportional frailty deficit within the study cohort.

7.2.1.3 The Distribution of Duration

The distribution of duration is an interval of the first to the last deficits recorded of each case within the study duration. The average duration in years for fit category [male: 3.6, female: 3.4], mild category [male: 8.1, female: 8.1], moderate [male: 10.2, female:10.3] and severe category [male: 11.6, female: 11.8] respectively. The boxplot in Figure 7.5 shows the distribution of duration in each frailty categories. It very useful in displaying the dispersion of datasets within different categories based on the five numbers summary (by referring to the bar in the figure, from bottom to top); the minimum (lower whisker), first quartile, median/middle quartile, third quartile and the maximum (upper whisker). The four sections separating by the five summary numbers are containing approximately 25% of the data in the dataset.

From the figure, it demonstrates that none of the categories follow normal distribution or having similar variability of duration in the dataset. Although both fit and mild categories distribution are right-skewed, there is difference between fit and the other three categories. It explained by the average duration by the fit category as it recorded with the shortest duration among another category with median [male: 1.9, female: 1.7]. However, the differences average of the other categories is not significant for median; mild [male: 7.9, female: 7.9], moderate [male: 10.6, female: 11.0] and severe [male: 12.2, female: 12.6].

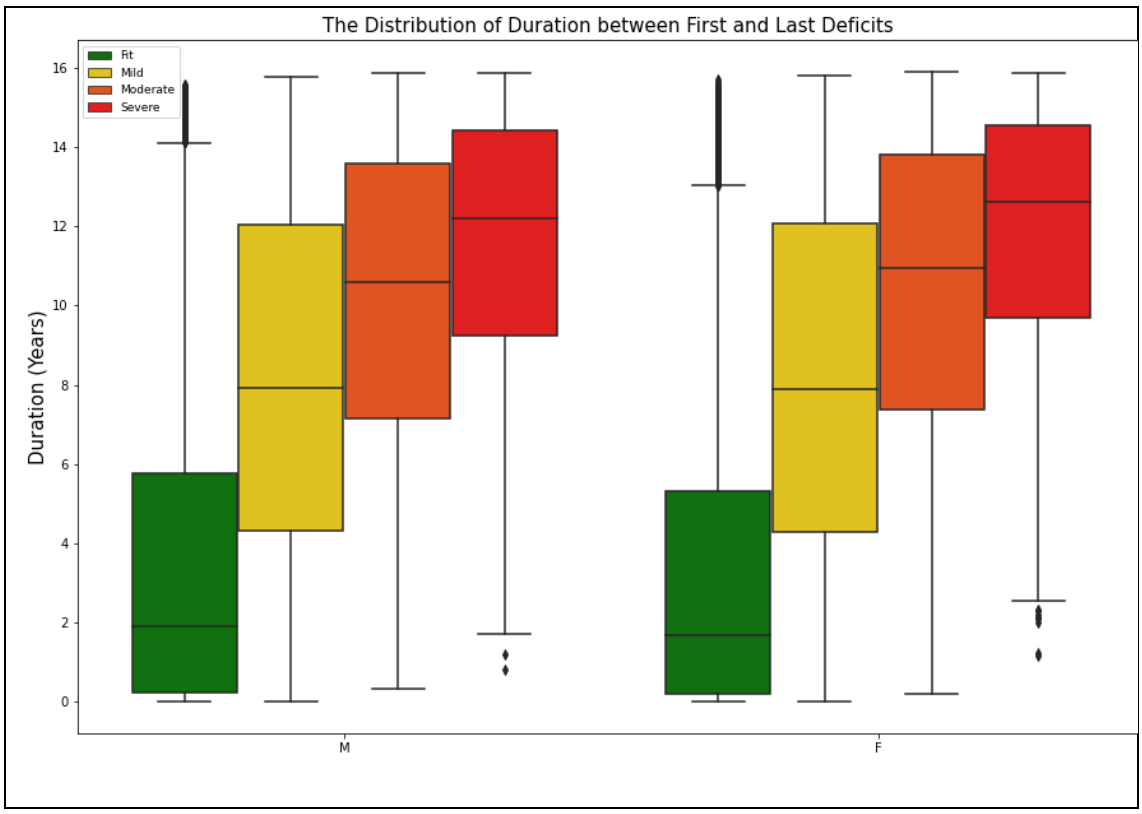


Figure 7. 5 The boxplot of the duration of patient case within the study duration. The x-axis indicates the duration of case interval in years, whereas the y-axis represents the four frailty categories.

7.2.1.4 The Distribution of Healthcare Professional Contact

Figure 7.6 presents the distribution of contact with health professionals within different frailty categories. The median and average showing that male cohort dominating in all categories compared to female (average), fit [male: 38(52), female: 33(47)], mild [male: 111(121), female: 104(113)], moderate [male: 177(188), female: 167(180)] and severe [male: 259(274), female: 244(266)]. The trends illustrates that the number of contact increases with the severity of frailty categories. However, the trend in standard deviation is showing differently as in the severe category female is leading the value with 119 and male 114.

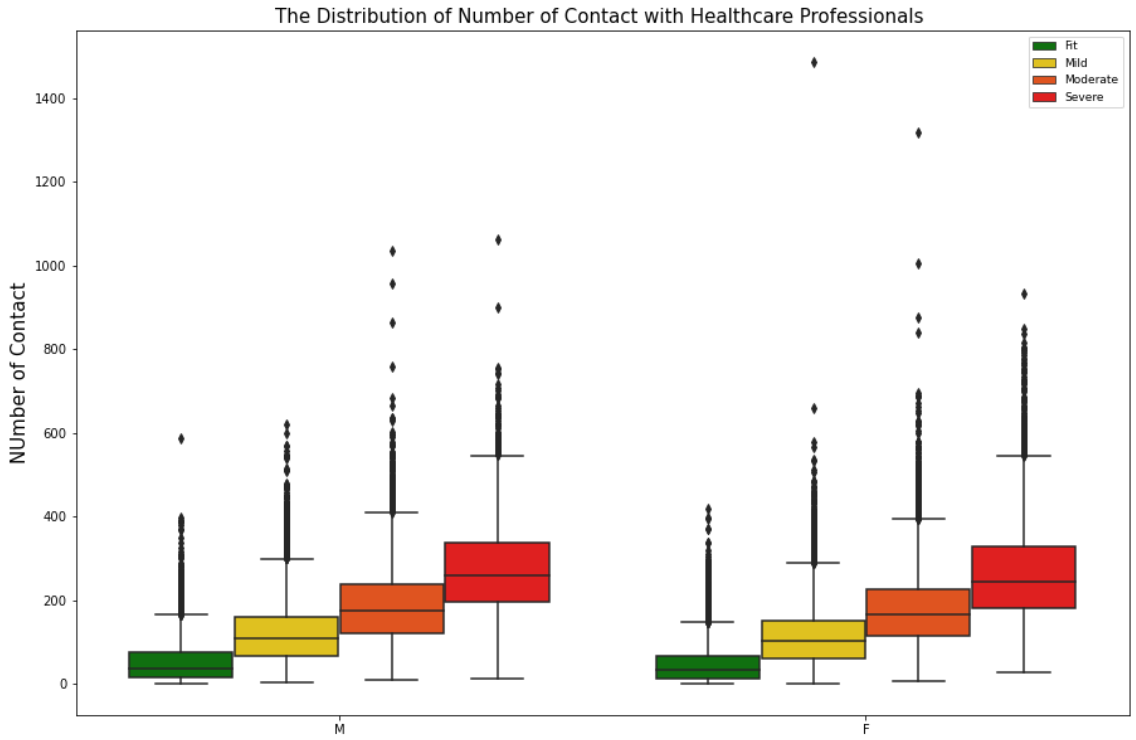


Figure 7. 6 The boxplot of the number of contact patient has with healthcare profesionals within the study duration. The x-axis indicates the duration of case interval in years, whereas the y-axis represents the four frailty categories.

7.2.1.5 Distribution of Events

The last distribution is the number of events (redundant deficits) within different frailty categories in both genders of male and female was presented in Figure 7.7. it shows that male is dominating the mean and median value in all categories with and increasing trend from fit category until severe. The fit category constitutes, mean(median) [male: 10.9(7.0), female: 10.2(6.0)], mild [male: 24.1(18.0), female: 22.9(17.0)], moderate [male: 41.9(35), female: 39.2(32.0)] and severe [male: 67.7(60.0), female: 64.6(55.0)].

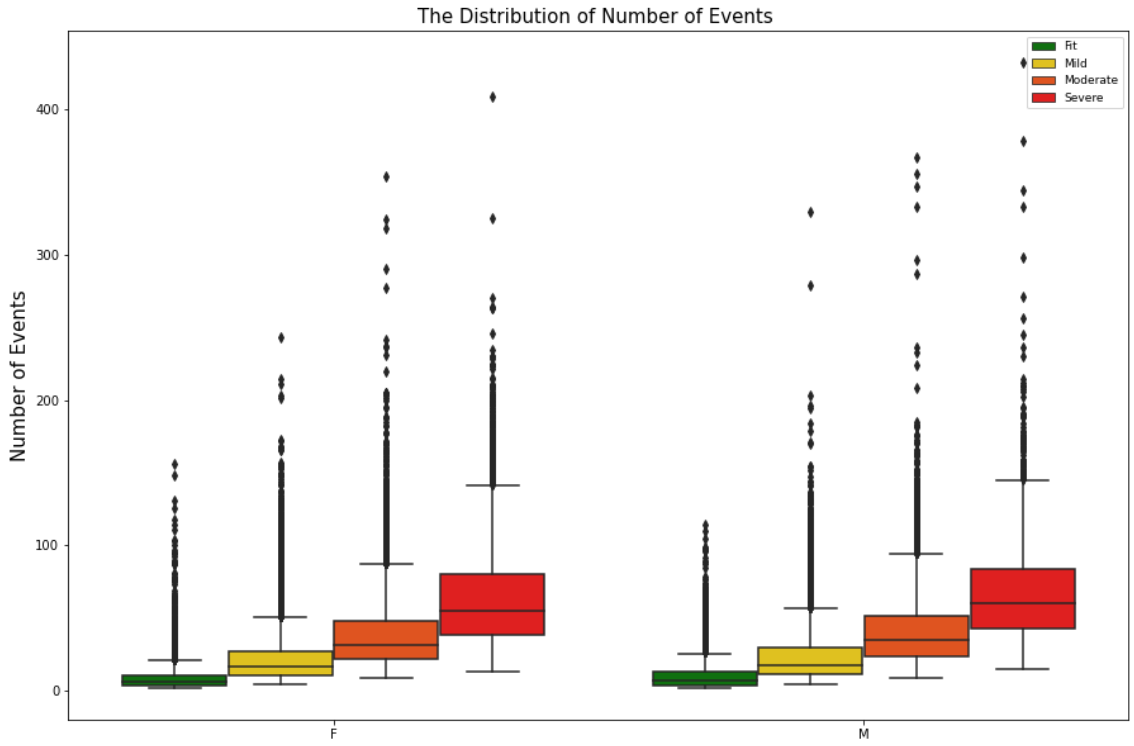


Figure 7. 7 The boxplot of the duration of patient case within the study duration. The x-axis indicates the duration of case interval in years, whereas the y-axis represents the four frailty categories.

7.3 Experiment 6: Cut Off Frailty Score Analysis

This experiment is done to determine the cut off points of frailty score. While in past study the identification of cut-off point followed the quantile based with point of 0.25 as the split points and the upper limit is 99th centile (Clegg *et al.*, 2016). However, in this work the cut off points is based on the feature variables in dataset. The area in between the subsequent split points will represent different frailty categories. The experiment follows the methodology described in Chapter 3.

7.3.1 Stage I: Planning

The first stage of the experiment involves identifying the goal and main research question. It aims in identifying the split points for different frailty categories consisting of not frail, mild, moderate, and severe. The primary research question is “Can feature variables of the dataset used to facilitate in identifying the split points of frailty scores?”. The question addresses the main questions of this research (RQ5) “Are the cut-off points used in electronic frailty index (eFi)

literature confirmed by real life data?", (RQ6) "Is it possible to determine the new cut-off points following data-based approach?" and (RQ7) "What feature can be used to characterize the new cut-off points in eFi scores?". This experiment examines a machine learning approach in answering the three main research questions by determining the feature variables available in the dataset. This experiment will set as a baseline in acquiring the cut off points to classify different frailty categories based on the feature variables of dataset.

7.3.2 Stage II: Extraction

The second stage is extraction is done by python query (notebook Jupyter) to obtain the dataset including the feature and target variables. The frailty score is feature variable with range of 0.03 until 1.00. Whereas the target variables are age at final deficit (based on unique deficits and based on redundant deficits count), duration between first and last deficits (based on unique and redundant deficits count), number of different healthcare professional contact, total number of healthcare professional contact, number of deficits (based on unique and redundant count) and number of events within the study duration.

7.3.3 Stage III: Transformation and Loading

The data transformation is done in third stage before loading into any analysis tool. Two steps involved in this stage are target variables selection and identifying the strength of association between the target and each feature variables. Classification of no strength of association will be removed. Other variation of strength of classification is included and will be discussed in detail in the next sub-section.

7.3.3.1 Measurement of Association

The distribution of individual variables was presented in previous section. In this section the association or correlation between two variables is explored. The first step is selecting the individual variables appropriate for the confirmatory analysis. Five variables which includes age at final deficit (based on unique deficit point), duration between first and last deficits (based on unique deficits point), number of GP contact, number of events and number of unique deficits are selected for the analysis. The variables which redundantly present within the study duration are filtered out from the analysis. The justification behind the selection step is the first unique deficits demonstrate a hallmark as to how the frailty progress.

The association between variables was examined between the features and target variables. Number of unique deficits identified as the feature variable while the other four variables denoted as target variables. It has reported in the past study the investigation of the frailty score cut off points need to be gender specific as risk of mortality responds to different level of frailty score between different gender (Hoogendijk *et al.*, 2020). Hence, association between target and each feature variables are measured within the elderly group. The correlation coefficient of statistical analysis is measured to identify the association between variables using Spearman's Rank Correlation Test as discussed in Chapter 3 in Section 3.4.2.

The correlation test is conducted between the number of unique deficits and four variables (age at final unique deficit, duration between first and last of unique deficits, number of contacts with healthcare professionals and number of events). The hypothesis to test is there is no correlation between the number of deficits and the other four variables within different gender of patient. The heatmap is presented to illustrates the strength of association between the each of variables and the number of deficits in Figure 7.8. Both axes in the figure represents the variables which has been simplified to deficit (number of unique deficits), age (age at final unique deficit), duration (duration between first and last of unique deficits), contact (number of contacts with healthcare professionals) and events (number of events).

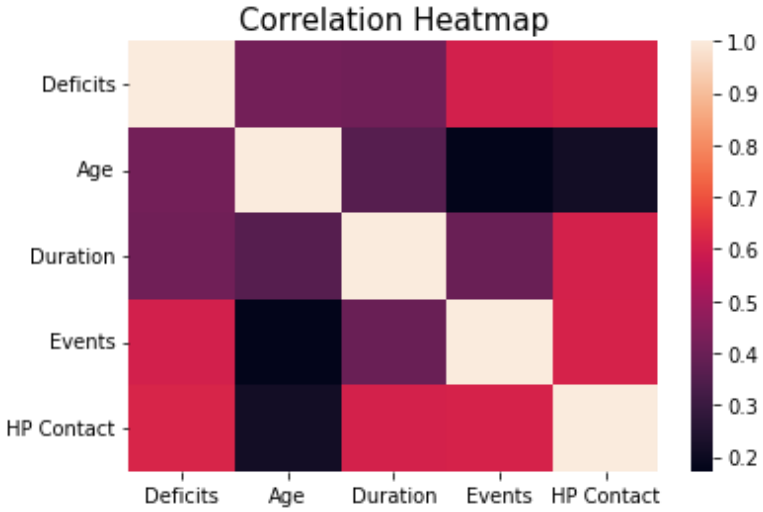


Figure 7. 8 The side-by-side heatmaps of different gender for Spearman's rank correlation value

Heatmap shows the association between deficits and four variables. It is a focus in this step as the selection of variables based on the strength of the association. The numerical information strength of association between each variable as

shown in Table 7.1. It is essential in determining the variables that has any association with the number of deficits.

Table 7. 1 Spearman's rank coefficient and p-value with the target variable, where the IoC represents as Interpretation of Correlation. The highest strength is highlighted

IoC	**High**	**Moderate**	**Low**	
Deficit	Event	**Contact**	Age	Duration
Spearman Coefficient	0.605	**0.618**	0.420	0.411
p-value	0.00	0.00	0.00	0.00

The null hypothesis is rejected in spearman's rank test of elderly patients which hypothesize that no correlation is present between number of deficits and all four variables. Is it observed that the highest positive correlation exists between deficit and contact with value of 0.618. Although two variables (age and duration) have low correlation value, they are still considered as feature variables in determining the cut off points of frailty scores. The reason is because the correlation value is more than 0.3 to be interpreted as negligible correlation (it is still correlated but with the smallest value).

7.3.4 Stage IV: Analysis

In this experiment, a quantitative analysis is done to determine the cut-off point of frailty score. It was done through the implementation of the classification technique following the discretization approach of optimal binning method as explained in Chapter 3 in Section 3.4.1.

The output of the PCA is presented in Figure 7.9 and portrays as distribution between the principal components and number of deficits. The principle component comprises of age of patient at final frailty deficit, duration between the first and last deficits, number of contact with healthcare professionals and number of events within the study duration.

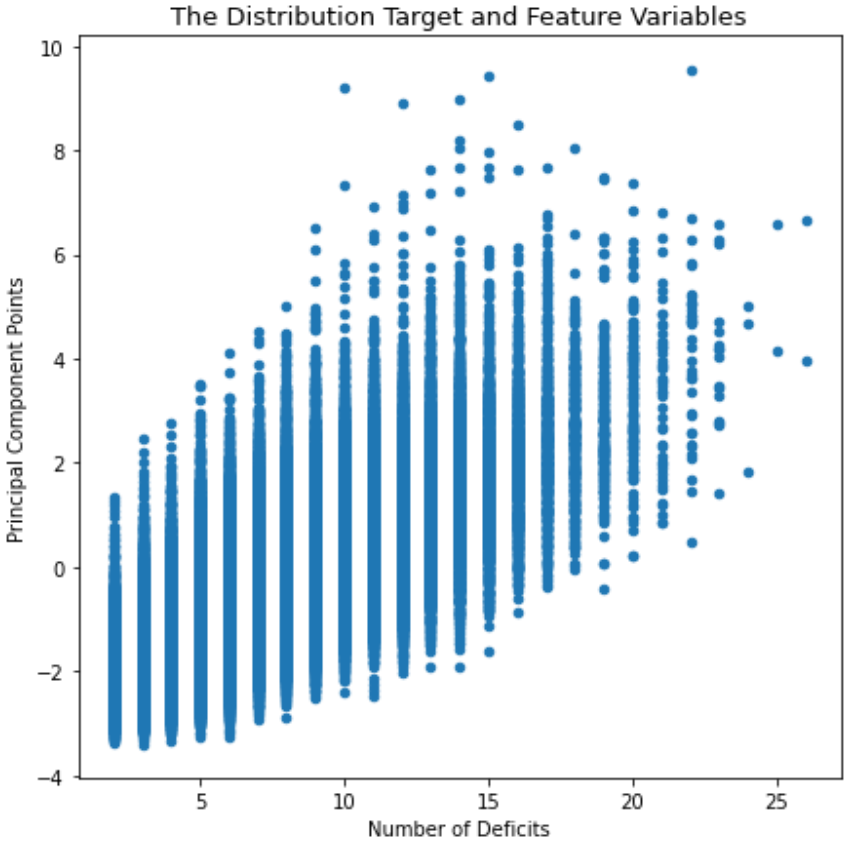


Figure 7. 9 The dataset distribution after employing the Principle Component Analysis

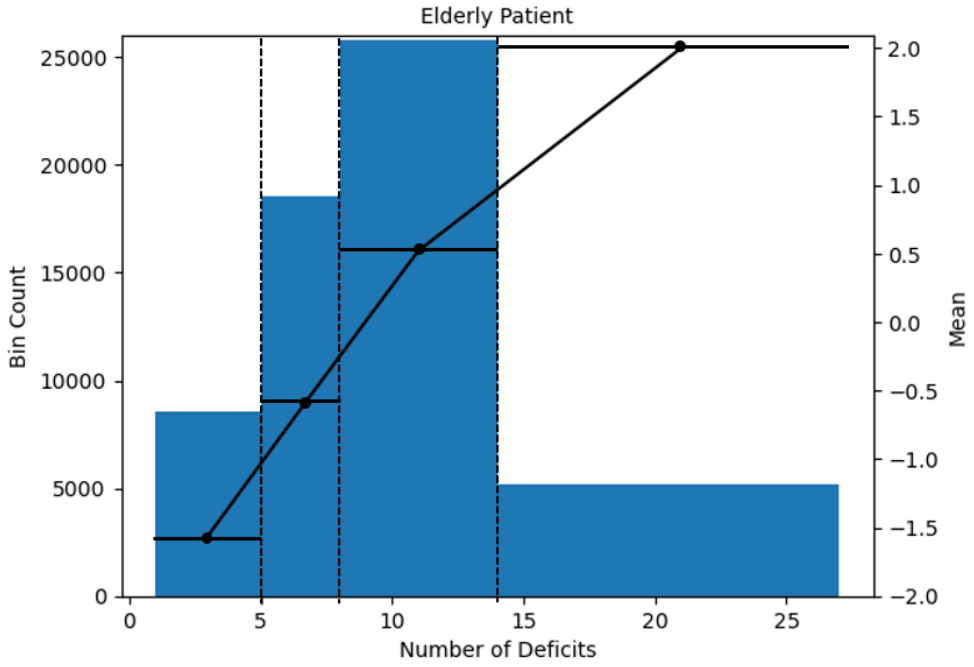


Figure 7. 10 The binning plot with the actual width of bins. The dotted lines represent the cut off points between bin or frailty categories

Figure 7.10 shows the result of discretization within male and female elderly. It illustrates the distribution of patients in each bin with their mean of principle component. The bar plot represents the number of patients in each bin while the dot is mean of principle component values. The widest width is at the last bin which consists of thirteen deficits. The mean principle component value for the first bin is -1.55, second is -0.55, third bin is 0.54 and last bin is 2.01. Table 7.2 shows the concise statistics of the splitting points between male and female with their related information. The cut off points of elderly is similar at the first two frailty categories of fit (not frail) and mild. However, female elderly has one extra deficit (with total of three) in the moderate category compared to male. As a result, the transition to the next frailty category is dissimilar making male reaching the final frailty category earlier than female.

Table 7. 2 The result of cut off points

Split Points	**[5, 8, 13]**			
Frailty Category	Fit	Mild	Moderate	Severe
Range of Deficits	1 – 4	5 – 7	8 – 12	13 – 36
# Deficits	4	3	5	34
# Patient	8,550	18,600	23,496	7,499
%	14.7	32.0	40.4	12.9

7.3.5 Stage V: Evaluation

The aim of the experiment is to determine the cut off points for frailty deficits within the elderly patient. The discretization was able to split the deficits value by setting up the maximum limit of bins with high measure of divergence between frailty categories. The measurement was based on the association between the feature and target variables of duration between first and last unique deficits, age at final deficit, number of events and number of contacts made with healthcare professional. The discretization result is summarised in Table 7.2. Although mild category having the least bin width of only three deficits, the proportion of patient is second highest of the cohort. While severe category with highest bin width has the lowest number of frail elderly.

This experiment gives an insight that there is potential in using association between attributes and number of deficits within the elderly patients. The work is a refinement approach in determining the cut off points derived from the target variables. The proposed cut off points will be further explored for comparison analysis with the domain expert cut off points or Clegg's approach.

The feedbacks from domain expert found the approach in this work in identifying the cut-off points for elderly stratification and is beneficial. It has proven to provide

an opportunity in exploring other data-based technique in determining the category based on available data within the healthcare system. However, other features variables with strong association is highly suggested to be included as an improvement to the future work.

7.4 Experiments with Proposed Cut off Points

An experiment will be conducted to model the frailty progression with the proposed cut off points. In addition, a comparison investigation is done as part of the confirmatory analysis with the cut off points reported in the literature (Clegg *et al.*, 2016). It starts with an exploratory data analysis based the proposed cut off points as an initial investigation of the frailty categories distribution. The experiments will follow the methodology explained in Chapter 3.

7.4.1 Exploratory Data Analysis of Cohort

The exploratory analysis will focus on the five attributes selected for the discretization approach. The pattern of distribution and numerical statistics are explored between different frailty categories and four attributes. The four attributes are duration between first and last unique deficits, age at last deficit, number of events and number of contacts with the healthcare professionals.

The proportion of elderly patient in each frailty categories is 44.43% in moderate, 31.99% in mild, 14.69% in fit and 8.89% in severe category as shown in Table 7.2. Figure 7.11 presents the age distribution of elderly patients within different frailty categories and genders. It shows the proportion of female patients is higher than male patients in most of age group, starting from [75 – 79] until [105 – 109]. Age group of [80 – 84] constitutes the largest number of patients (in both genders) where moderate as the dominating category. The pattern follows by the age group of [85 – 89] of which zero patient in fit category.

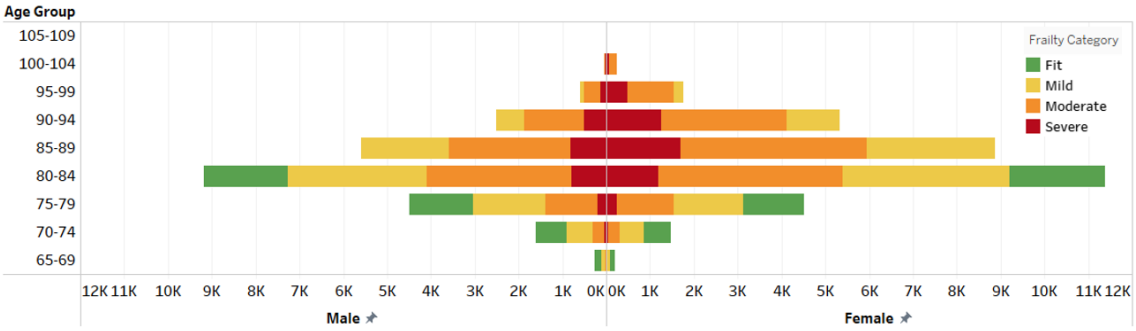


Figure 7. 11 The pyramid population of elderly patients. The y-axis shows age range (five years) and x-axis represents the number of patient in respective age range. The different colour to differentiate the frailty categories.

The duration is an interval in years between the first and last unique deficits within the study duration in each case. The average of case duration of cohort is 9.4 years with median of 9.8 years. While across gender, the average case duration is slightly higher in female (9.6 years) compared to male (9.1 years) as well as the median (female 10 years and male 9.3 years). Figure 7.12 shows the distribution of statistical information of duration within different frailty categories and gender using the boxplot visualisation. It is a useful visualisation in showing an individual numerical information across multiple categories. The plot is based on five number of summary (referring to a boxplot from bottom to top) the minimum (lower whisker), first quartile, median/middle quartile, third quartile and the maximum (upper whisker). The four sections separating by the five summary numbers are containing approximately 25% of the data in the dataset.

It demonstrates that median average duration is increasing from fit to severe categories in both genders. The shortest is in fit category (female 5.2 years, male 5.1 years) and the longest is in severe category (female 13.0 years and male 12.5 years).

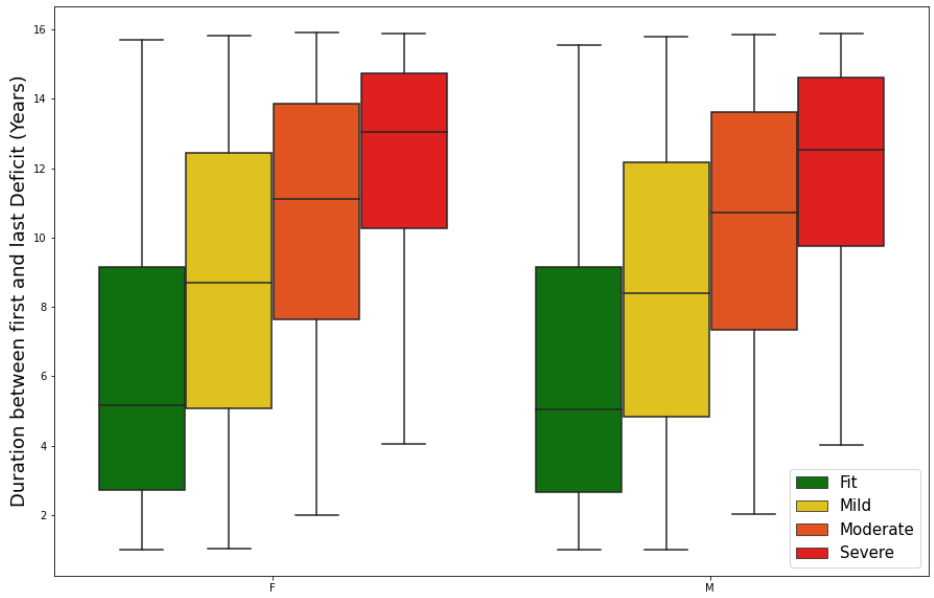


Figure 7. 12 The boxplot of duration in years between first and last unique deficits. The x-axis indicates the duration in years while y-axis of character 'F' represents Female and 'M' Male.

The second boxplot in Figure 7.13 shows the distribution of number of contacts with healthcare professionals within different frailty categories between female and male. Male is dominating the number of contacts in each frailty categories compared to female. The highest average number of contacts made is in male from severe frail category 292 (median 277) while female from severe frail category is 287 (median 265). It follows by moderate [male: average 188, median 175], [female: average 181, median 168], mild [male: average 121, median 113], [female: average 117, median 109], and fit [male: average 77, median 68], [female: average 77, median 68] categories.

The last boxplot is distribution of number of events within different frailty categories of female and male elderly in Figure 7.14. Male is still dominating the total number of events recorded within the study duration where the highest is in severe category with average 72 (median 64) while female average is 70 (median 61). The second highest is in moderate category with [male: average 41, median 33], [female: average 38, median 31], follows by mild [male: average 22, median 16], [female: average 21, median 15] and fit category [male: average 11, median 7] and [female: average 11, median 6].

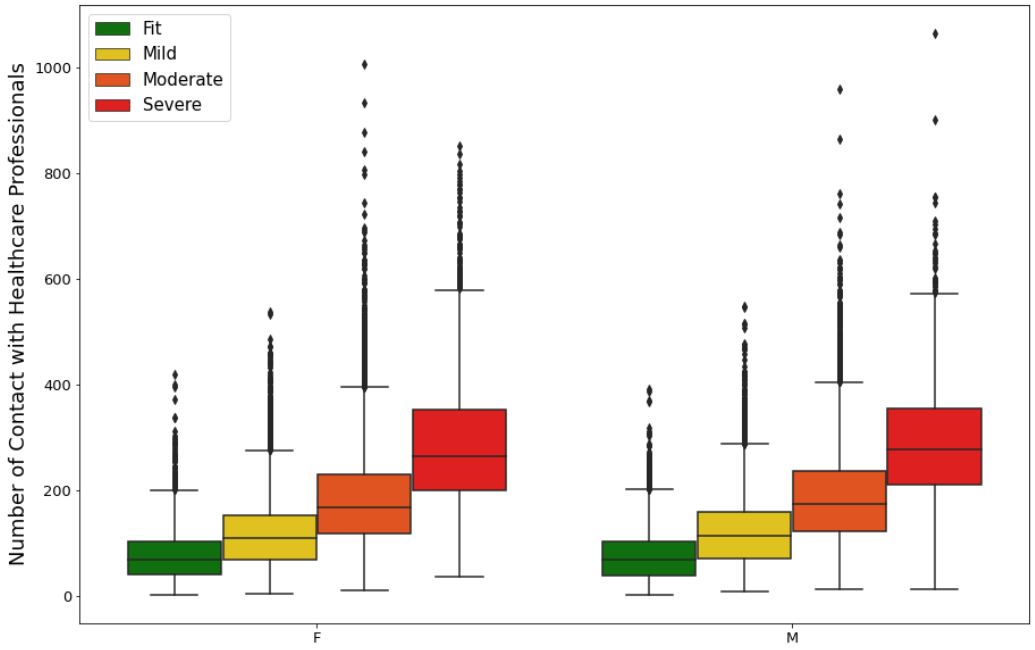


Figure 7. 13 The boxplot of number of contacts within different frailty categories and gender.

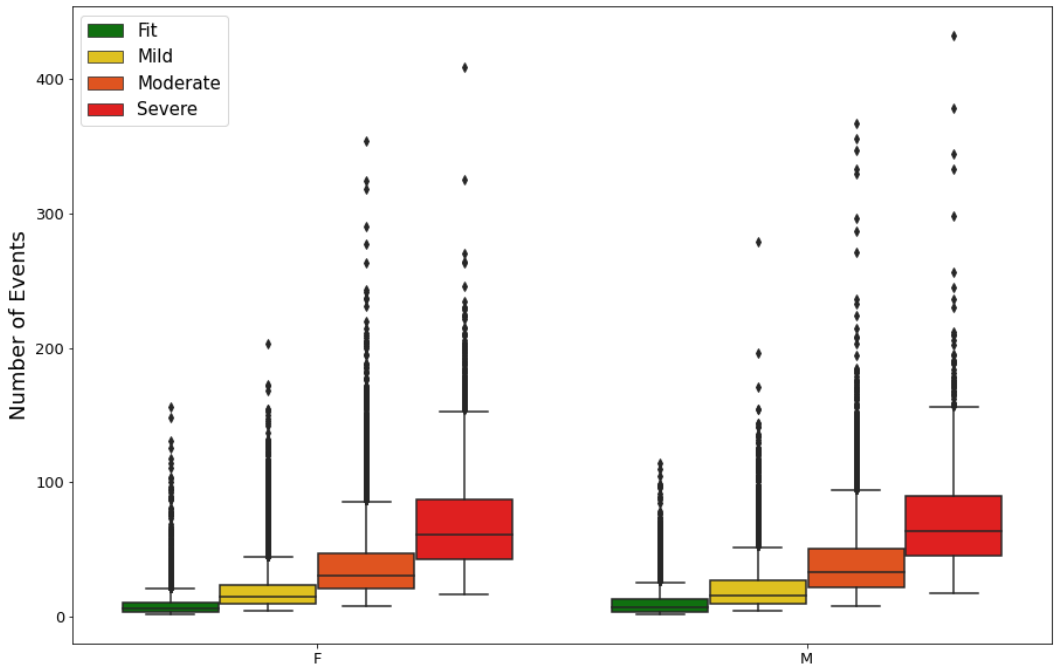


Figure 7. 14 The boxplot of number of events within different frailty categories and gender

7.4.2 Experiment 7: Comparative Analysis between Cut Off Points used in Literature and Proposed Cut Off Points

This experiment was done to compare the application of proposed cut-off points with the literature cut-off points, known as Clegg's cut-off points (Clegg *et al.*, 2016). The comparative analysis was conducted using similar dataset in determining the severity of frailty progression. It aims to reveal the diverse pattern of frailty progression at each stage based on different cut-off points applied. This comparative analysis is an extended version of analysis in Section 6.4. The proposed cut-off points used was derived in the previous Section 7.3 and follows the general methodology discussed in Chapter 3. To avoid repetition, only the extended step will be explained further in this section.

7.4.2.1 I: Planning

This study expands the analysis of frailty progression by examining the frailty progression within different classification of elderly. The classification of elderly describes based on the 10-years range starting from 65 years. The first age group of elderly is (65 – 75) known as youngest-old, (75 – 84) is middle-aged elderly and 85 years and above is oldest-old elderly. The aim of this experiment is to answer the primary research question "What are the differences in frailty progression and its association with focused deficits between the cut-off points used in literature and the proposed cut-off points?". A comparison-based analysis between the cut-off points available in past studies and proposed cut-off points in previous experiment 6 in Section 7.3 is established. It relates to the main research question of (RQ5) compare with the proposed cut-points in previous experiment, (RQ6)-using data-based approach utilizing the discretization technique and (RQ7)-using the feature variable identified in experiment 6 respectively. The research question in this experiment was addressed by investigating the frailty progression within frailty stages from fit, mild, moderate and to severe among different elderly classification. It also examines frailty progression in association with deficits of concern such as fall, hypertension and polypharmacy.

7.4.2.2 Stage II: Extraction

The extraction was performed from Bradford SQL Server database management system. The main criteria for patient records extraction were elderly aged above 65 years within the study duration. The dataset was converted into raw event log that consists of *patient id, activity, timestamps,* and *resources details*. *Patient id*

hold unique number to determine every records that belong to particular patient. The *activity* was obtained mainly from the `ClinicalData` in Table 5.6 with the name of activity was based on disease (e.g., heart disease, hypertension) and health status of the patient (e.g., housebound, difficulty in moving). The timestamp which belongs to each activity is used directly in date format.

Additional selection criteria were applied to ensure the dataset selected is fit for the analysis. It includes (1) patient records within study duration should be at least one year, (2) the final frailty category for patient aged 85 years above need to be in either moderate or severe stage, (3) maximum annual accumulation of frailty deficits is three and final inclusion is (4) patient with combination of all three deficits of concern: fall, hypertension, and polypharmacy. This stage has extracted 8,547 elderly patients.

7.4.2.3 Stage III: Transformation and Loading

Transformation is a stage to convert the extracted dataset into several event logs for process mining analysis. Three steps were applied which includes (1) selecting only the first occurrence of frailty deficits, (2) enriching the event log with additional events of frailty stages: fit, mild, moderate, and severe (3) securing the sequence of additional events within the primary events in the event log. The initial step considers only the first occurrence of deficit to target the newly emerged frailty deficits in the progression of frailty. The second step was done in two separate event logs in determining the frailty categories, (i) following the Clegg's approach and (ii) following the proposed cut-off points based on the previous experiment in Section 7.3.

Table 7. 3 The difference of cut off points between two approaches

Frailty Stages	Range of Accumulated Deficits			
	Clegg's	Count of Deficits	Proposed	Count of Deficits
Fit	1 – 4	4	1 – 4	4
Mild	5 – 8	4	5 – 7	3
Moderate	9 – 12	4	8 – 13	6
Severe	13 – 36	24	14 – 36	23

Table 7.3 presents the difference of cut-off points between Clegg's approach and this study proposed approach. It shows the minimum accumulated deficits required to reach certain frailty stages, for example in Clegg's approach a patient who has nine (9) is in moderate stage, while in the proposed approach the minimum accumulated deficits of eight (8) is already in the moderate frailty stage. Later, the identified frailty stages as the additional events were combined into the

primary events. The sequence of the combined log was rearranged to only permit fit stage event to be the first event in all cases and other additional events at the end of the primary events. Example arrangement of other additional events such that for Clegg's approach is mild event was positioned after unique deficit number four (4), moderate after eight (8) and severe after twelve (12). The rationale of this method is that the arrangement explained above is to confirm patient frailty stage indicated by the minimum number of deficits acquired.

The two transformed event logs were split based on the age classification of (65 – 74), (75 – 84) and (85+). The splitting step was done through the Python Jupyter tools based on patient age at their first frailty deficit in the record. The split event log was produced in the format of .CSV file which was loaded into Disco and ProM tools for mining stage.

7.4.2.4 Stage IV: Mining and Analysis

Process mining techniques of process discovery and performance checking were used apart from, two perspective of process mining which are control flow case-based analysis and time perspective in mining and analysis stage. The quantitative measures were conducted to examines the frailty progression through process variant analysis.

Two event logs were created from a similar dataset following two approaches of cut-off points, Clegg's, and the proposed approaches. The logs resulted a total of 8,547 number of patients, 30,754 events with approximately 4 events per patient. The interval between the start and end of each frailty stages and transition points were determined from the two logs and shown in Table 7.4. The transition points are defined as the transition of frailty progressing into more advance frailty state such as from end of fit to start of mild (transition point 1), from end of mild to start of moderate, (transition point 2) and from end of moderate to start of severe (transition point 3). In general, the interval between stages and transition points were similar whereas, the only difference observed were in the frailty stages of mild and moderate. The difference is Clegg's approach took shorter time in moderate stage (median duration of 1.6 years) while the proposed approach had shorter duration in mild stage (with median duration 1.4 years).

Table 7. 4 The interval within frailty stages and transition points of two set approaches

Frailty Stages	Clegg's Cut-Off Points	Proposed Cut-Off Points
	Median Duration (IQR) In Years	
Fit	2.2 (1.1 – 3.9) N = 8,547	2.2 (1.1 - 3.9) N = 8,547
Mild	**2.3 (1.2 - 4.0)**	**1.4 (0.6 - 2.7)**

	N = 8,514	**N = 8,514**
Moderate	**1.6 (0.5 - 3.1)** **N = 7,023**	**2.3 (1.0 – 4.2)** **N = 7,025**
Severe	0.9 (0.0 – 2.8) N = 3,335	0.9 (0.0 – 2.8) N = 3,335
Transition Point	**Median Duration (IQR) In Years**	
1	0.6 (0.2 – 1.4) N = 8,514	0.6 (0.2 – 1.4) N = 8,514
2	0.6 (0.2 – 1.4) N = 7,023	0.5 (0.2 – 1.4) N = 7,025
3	0.5 (0.2 – 1.2) N = 3,335	0.5 (0.1 – 1.2) N = 3,335

Both cut-off points approach was split based on age classification (sub-cohorts) and its descriptive measures is shown in Table 7.5. This step aims to show the variation in interval between stages and transition points within sub-cohorts. It observed that interval within the mild stage in Clegg's approach is longer than moderate stage in all sub-cohorts while in the proposed approach is moderate is longer than mild stage and consistent in all sub-cohorts. The shorter interval is moderate stage (compared to mild stage) in all sub-cohort following Clegg's approach: in 65 – 74 (median interval of 1.8), 75 – 84 (median interval 1.6) and in 85++ (median interval of 1.1). In contrast, following the proposed approach mild stage experienced the shorter interval compared to moderate in 65 – 74 (median interval of 1.7), 75 – 84 (median of 1.4) and 85++ (median of 1.0).

Table 7. 5 The numerical information of interval, median (mean) within frailty stages and transition points. The highlighted row represents the difference between the approaches of two cut-off points

	Clegg's Cut-off Points			**Proposed Cut-off Points**		
	65 – 74	75 – 84	85++	75 – 84	65 – 74	85++
Fit Stage	2.7 (3.3)	2.1 (2.7)	1.8 (2.2)	2.7 (3.3)	2.1 (2.7)	1.8 (2.2)
Transition Point 1	0.7 (1.2)	0.6(4.1)	0.4 (3.3)	0.7 (1.2)	0.6 (4.1)	0.4 (3.3)
Mild Stage	**2.7 (3.2)**	**2.2 (2.7)**	**1.6 (2.1)**	**1.7 (2.2)**	**1.4 (1.8)**	**1.0 (1.4)**
Transition Point 2	**0.7 (1.1)**	**0.5 (0.9)**	**0.4 (0.8)**	**0.7 (1.1)**	**0.6 (1.0)**	**0.4 (0.7)**
Moderate Stage	**1.8 (2.3)**	**1.6 (2.0)**	**1.1 (1.5)**	**2.7 (3.1)**	**2.2 (2.7)**	**1.8 (2.2)**
Transition Point 3	0.5 (0.9)	0.5 (0.9)	0.4 (0.8)	0.5 (0.9)	0.5 (0.8)	0.4 (0.8)
Severe Stage	1.1 (1.9)	0.9 (1.8)	0.4 (1.2)	1.1 (1.9)	0.9 (1.8)	0.4 (1.2)

The next analysis was done by examining the pattern of sequence (PoS) in each sub-cohort with the deficits of concern such as fall, hypertension and polypharmacy. The PoS identified for both cut-off points approaches were like in Figure 6.4 in Section 6.4 as similar dataset was used. The PoS I is dominant where fall occurred after both hypertension and polypharmacy. It followed in the (65 -74) and (75 – 84) age sub-cohorts as shown in highlighted cell of Table (H) in Appendix A.7 of experiment documentation. Table 7.6 presents the descriptive statistics of trace variants in sub-cohort following the PoS I in sub-cohort (65 – 74) and (75 – 84). It shows the sub-cohorts in both age classification has similar

numerical measures of mean, median case duration but with slightly different in number of trace variant, in (75 – 84) of Clegg's approach with 64 variants while in proposed approach with 61 variants. The results based on the descriptive measures of log variant revealed no obvious differences between the sub-cohorts.

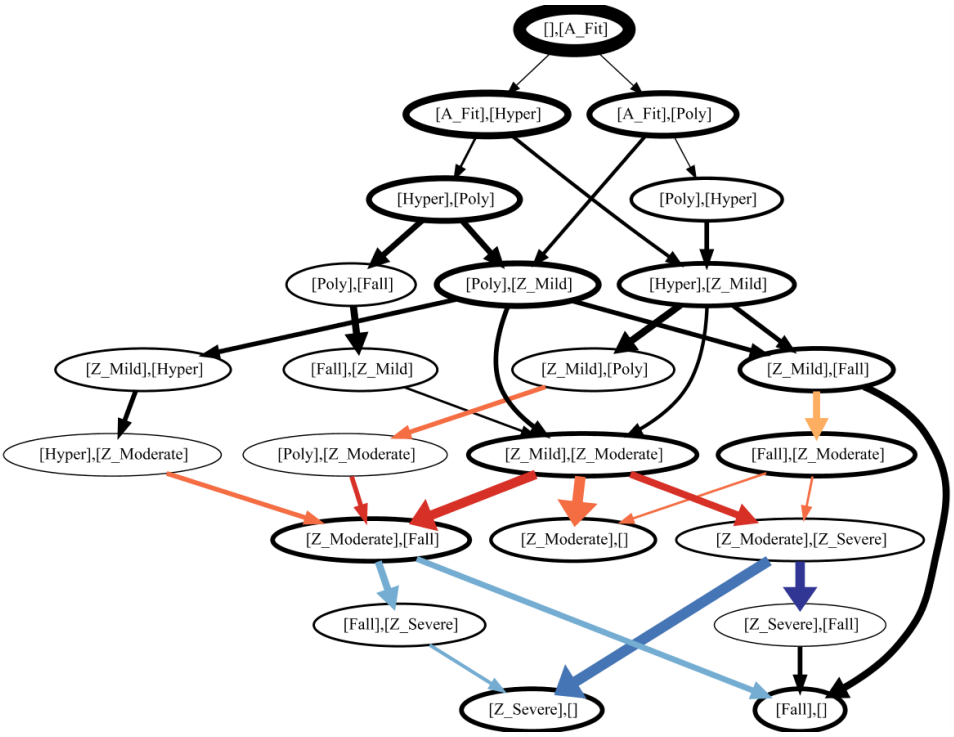
Table 7. 6 The descriptive information of the dominant PoS I with the mean and median case duration in years

Cut-off Points Approach	Clegg's	Proposed
	(65 – 74)	
# Patient (%)	2,266 (26.5%)	
# Events (%)	14,198	14,115
# Trace Variant	61	61
Case Duration, Mean [Min- Max] (days)	2.0 [0.0 – 14.1]	2.0 [0.0 – 14.1]
Case Duration, Median [IQR] (Days)	1.2 [0.1 – 3.1]	1.1 [0.1 – 3.0]
	(75 – 84)	
# Patient (%)	2,566 (30.0%)	
# Events (%)	16,061	15,980
# Trace Variant	64	61
Case Duration, Mean [Min- Max] (days)	1.7 [0.0 – 13.0]	1.7 [0.0 – 13.0]
Case Duration, Median [IQR] (Days)	1.0 [0.1 – 2.5]	1.0 [0.1 – 2.5]

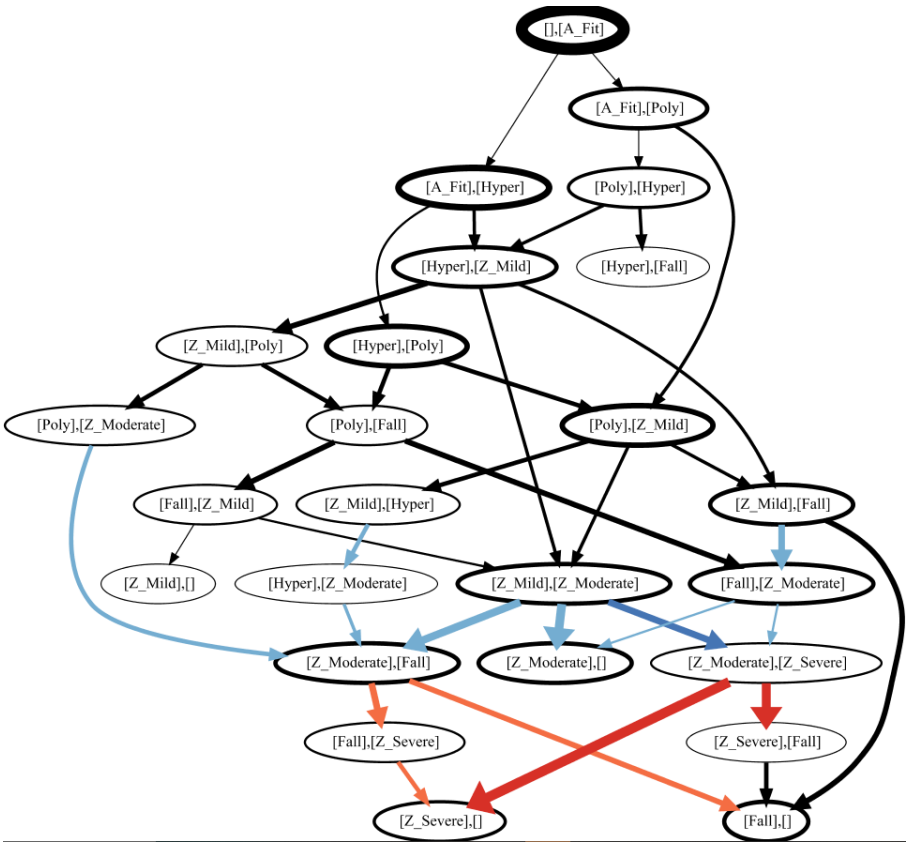
Although, the descriptive statistics illustrates similar measures between the PoS I in both sub-cohort of (65 – 74) and (75 – 84), the significant pattern of association between the deficits of concern within the different approach of cut-off points is unknown. Hence, the next comparison was done using the *Process Comparator* (Bolt, de Leoni and van der Aalst, 2018) plug-in in ProM. The plug-in determines the significant differences between different variants in the process. Performance comparison was performed between the states label in comparing the time differences of activities execution. The widely used confidence level of alpha value of 0.05 was applied throughout in the analysis.

The Process Comparator plug-in produced a transition system and analysis based on the abstraction of the last two events were assessed. The duration of execution between activities is the main process metrics being evaluated. The transition system in comparing the cut-off points approach was shown in Figure 7.15 of (a) sub-cohort (65 – 74) and (b) sub-cohort (75 – 84). The statistically significant transition is shown by coloured arrows. The blue shades colour represents the higher metric (significance difference or effect size) in sub-cohort following Clegg's approach while red shades represent higher metric in the proposed event log as shown in colour legend in top figure of Figure 7.15. While the thickness of the arrow indicates higher duration of transition between state labels compared to other arrows in the transition system.

Many of the relevant performance (duration of transition) detected in the lower half part of both transition system. The lower part of transition system in Figure 7.15 (a) and (b) demonstrates the stages in frailty progression from mild to severe. However, the difference can be observed is that the proposed approach has higher statistically significance in transition duration between state label in sub-cohort (78 – 84) between the moderate to severe frailty stage as shown in Figure (b) 7.15, while in sub-cohort (65 – 74) the proposed approach has higher statistically significance in duration transition between the mild and moderate frailty stage illustrates in Figure (a) 7.15. The opposite pattern was observed in Clegg's approach.



(a) Sub-cohort (65 – 74)



(b) Sub-cohort (75 – 84)

Figure 7. 15 The results from the Process Comparator Plug-in using two sub-cohort of (65 – 74) and (75 – 84)

The next step in examining the frailty progression was performed on the dominant transition between the sequence of state labels. The finding of the step is presented in evaluation section.

Stage V: Evaluation

The evaluation stage was conducted through the analytical assessment using finding from this experiment with the literature as part of the confirmatory analysis. In addition to that, the feasibility of the approach also been assessed in relation to the applicability and suitability in answering the research questions.

The analytical assessment in determining the differences of frailty progression between the cut-off points approach were started by identifying the dominant transition within the sequence of state labels. Table 7.7 presents the trace variant comparison of PoS I between the two cut-off points approach of (65 – 74) and (75 – 84) sub-cohorts. The highlighted value in Table 7.7 indicates the statistically significance differences of the two logs which trace frequency average is higher than the other log. The representation of the states label was illustrated following the abstraction of the last two events. For example, state label [`Mild`], [`Fall`] is

a single state with the last event `Mild` combined with current event `Fall`, while symbol > to indicate the transition to another state label [`Fall`], [`Moderate`] with the next event is `Moderate`. Hence, the transition described the progression of `Fall` to `Moderate`. The numerical information in Table 7.7 represents the average duration of the transition between states labels in years.

The behaviour between the cut-off points approaches was compared in the sub-cohort (65 – 74) and (75 – 84). In general, the dominant states label between the compared event logs of the two sub-cohorts were similar. The justification on the result is all compared sub-cohorts are from the similar PoS I. Most of the statistically significance difference of sequence of state label involved only deficit of concern, fall. This is in line with past studies as large volume of published literature studies indicating that fall is major outcome in frailty (Shim *et al.*, 2011; Curcio, Henao and Gomez, 2014; Kojima *et al.*, 2015; Fhon *et al.*, 2016; Cheng and Chang, 2017). It is apparent from Figure 7.15 that most statistically significance transition between states label which Fall was involved was between frailty stages of Moderate to Severe.

The pattern dissimilarity of the sequence of state label: [`Fall`], [`Moderate`] > [`Moderate`] is statistically significant higher in sub-cohort (65 – 74) for the proposed approach (24.2% vs 23.3%) with the standard deviation duration (0.9 vs 1.2). While the sub-cohort (75 – 84) the alike sequence state is statistically significance higher in Clegg's approach (24.8% vs 24.1%) with standard deviation duration (1.1 vs 0.7). Although, the transition duration (in standard deviation) was observed to be shorter in the proposed approach than Clegg's in both sub-cohorts of (65 – 74) and (75 – 84). The shorter duration in the proposed approach could be due to the lesser total count of deficits at Mild stage, three (3) deficits than Clegg's approach, four (4) deficits before proceeding to the Moderate stage.

Table 7. 7 The summary of states label and its transition with frequency average on either approach more than 10%. The numerical information presents the percentage of average frequency (average duration of interval, standard deviation of duration in years).

State Label	Clegg's	Proposed
(65 – 74)		
[Mild], [Fall] > [Fall, [Moderate]	**27.1 (3.4, 2.6)**	24.6 (2.9, 2.4)
[Mild], [Moderate] > [Moderate], [Fall]	16.8 (4.6, 2.5)	**25.5 (3.4, 2.2)**
[Moderate], [Fall] > [Fall]	11.4 (2.0, 2.0)	**24.8 (2.8, 2.5)**
[Fall], [Moderate] > [Moderate]	23.3 (1.2, 2.0)	**24.2 (0.9, 1.6)**
[Fall], [Severe] > [Severe]	**19.9 (1.2, 1.6)**	19.4 (1.8, 2.1)
[Moderate], [Fall] > [Fall], [Severe]	**18.5 (2.7, 2.2)**	16.6 (3.6, 2.5)
[Fall], [Moderate] > [Moderate], [Severe]	**11.7 (1.1, 1.5)**	5.5 (0.8, 1.2)
(75 – 84)		
[Mild], [Fall] > [Fall], [Moderate]	**26.5 (2.9, 2.5)**	23.7 (2.3, 2.1)
[Mild], [Moderate] > [Moderate], [Fall]	16.6 (3.8, 2.3)	**25.6 (3.0, 2.0)**
[Moderate] > [Moderate], [Fall]	11.9 (1.9, 1.8)	**25.3 (2.5, 2.2)**
[Fall], [Moderate] > [Moderate]	**24.8 (1.1, 1.6)**	24.1 (0.7, 1.3)
[Fall], [Severe] > [Severe]	**19.5 (1.2, 1.6)**	19.1 (1.9, 2.2)
[Moderate], [Fall] > [Fall], [Severe]	**18.2 (2.3, 1.9)**	16.6 (3.1, 2.4)
[Fall], [Moderate] > [Moderate], [Severe]	**11.4 (0.9, 1.3)**	5.7 (0.5, 1.0)

Apart from that, it is observed that the transition of sequence states label from Fall to Moderate is prominent in Clegg's approach: [`Mild`], [`Fall`] > [`Fall`], [`Moderate`] (27.1% with duration of standard deviation, 2.6) in (65 – 74) sub-cohort. While [`Mild`], [`Fall`] > [`Fall`], [`Moderate`] (26.5% with duration of standard deviation, 2.5), [`Fall`], [`Moderate`] > [`Moderate`] (24.8% with duration of standard deviation, 1.6) in (75 – 84) sub-cohort. Whereas the proposed approach shows significant pattern of sequence state label from Moderate to Fall. Two sequences of state labels with that pattern are (1) [`Mild`], [`Moderate`] > [`Moderate`], [`Fall`] in sub-cohort (65 – 74) (25.5% with duration of standard deviation 2.2), in sub-cohort (75 – 84) (25.6% with duration of standard deviation 2.0) and (2) [`Moderate`], [`Fall`] > [`Fall`] in sub-cohort (65 – 74) (24.8%, with duration of standard deviation 2.5) and sub-cohort (75 – 84) (25.3%, with duration of standard deviation 2.2).

This experiment explores an approach for comparison analysis on different cut-off points of frailty scores (eFI) on frailty progression. The association with deficits of concern with frailty progression was examined using process mining techniques. The domain expert found the approach demonstrates the potential of process mining in examining the frailty progression and its significance association between deficits of concern. They also reveal the visualization of process model using the Process Comparator plug-in illustrates the association of deficits of concern with frailty progression provides a straightforward understanding to other healthcare stakeholder.

In addition to that, it supports the comparability of two cut-off points in eFi scores to determine frailty stages. The Clegg's and proposed cut-off points exhibit slight

difference in general, relatively on the position of Fall deficit. However, it revealed that in Clegg's approach Fall is highly likely to progress the frailty progression to Moderate stage. While the proposed approach showed the opposite pattern, in which Fall occurred after elderly advancing into second last frailty stage, Moderate.

7.5 Summary

This chapter has presented two experiments using Bradford connected dataset. The first experiment investigated the new cut-off points (proposed cut-off points) based on the features available with the dataset. The features aim to determine the optimum cut-off points to differentiate different frailty stages. The second experiment demonstrated the comparison analysis between different cut-off points and the association of frailty progression with deficits of concern. It was an extended version of experiment done in the previous Chapter 6, experiment five (5). The approach and findings of the analysis can be observed in two viewpoints: frailty domain viewpoint and process mining viewpoint.

Initially, in frailty domain viewpoint the proposed approach in first experiment provides alternative in determining the cut-off points for frailty stages identification apart from the quantile method used in the Clegg's approach. The available features in the routinely collected healthcare data has shown that five features have enable the cutting-off points of frailty score to comprehensively segregate different frail elderly. On the other hand, the process models discovered in the form of transition system are helpful to visualise the frailty progression in experiment 7. Further analysis was done in identifying the significance difference of frailty progression between two cut-off points approaches was pertinent in medical domain.

Secondly, in process mining viewpoint the analysis has proven as evidence that routinely collected health record in Bradford dataset can be used in the work of process mining in frailty domain. In addition to that, combination of process mining and machine learning techniques to analyse frailty progression such as discretization-based analysis, trace variant analysis and pattern of event sequence identification resulted in helpful visualisation of frailty progression rather than a single-direct used of process mining technique. Furthermore, the transition state label in the second experiment (experiment 7) which has been uncovered using one of process mining techniques is very effective in revealing the significant direct transition between one frailty deficit to other deficits or frailty

stages within two cut-off points approaches. It is useful in understanding the different progression of frailty via two cut-off points.

The next chapter encapsulates discussion on previous Chapters 1 until 7 of this thesis. It provides as focus point of the thesis in exploring and criticizing the developed method for applying in the case studies on the experiments conducted in Chapter 4, 6 and 7.

Chapter 8

Discussion and Conclusion

Chapter 4, 6 and 7 present the analysis chapter of understanding and modelling the frailty progression based on the methodology described in Chapter 3. In this chapter discussion on the findings and lesson learned of the case studies using two datasets is presented. It includes the challenges of working with healthcare datasets using process mining and the work in modelling frailty progression. Finally, the confirmatory analysis between the cut off points reported from literature and proposed cut off points. At the end of this chapter, a contribution to the body of knowledge is summarized.

8.1 Challenges working with healthcare datasets

Several challenges were identified while working using real life healthcare data. The key challenges are getting the approval and access to the data, the quality of the data, understanding the healthcare data and deciding the best visualisation for results illustration. Each of the challenges were explained and the strategies taken to minimise the challenges were discussed as follow.

8.1.1 Approval to Data Access

The nature of the healthcare data is sensitive and private as it contains very personal health information. Hence, acquiring and attending to this kind of data require strict procedure. Furthermore, different organisation managing healthcare data generally has different and unique procedure in acquiring the data. Understanding the requirement and the rule of safety in accessing the data has becoming the challenge. In this study, two healthcare datasets were used with two varied approaches taken to access the data.

The first data is MIMIC-III database which is an openly available healthcare dataset. The access is secured after fulfilling the National Institute of Health (NIH) course as first requirement. The training is an online based course training which took a day to complete. The data requester needs to upload the certificate upon completing the training together with the MIMIC-III access request over the login PhysioNetWorks account. The certificate of completing the training course as illustrated in Appendix A.1. The requester needs to sign data user agreement as the second requirement. When all requirements have fulfilled, MIMIC-III dataset

can be downloaded in the type of .CSV files. The dataset comprises of twenty-six (26) different .CSV files which can be imported into any available database framework. The PostgreSQL was used in this study to manage the acquired data.

The second dataset is the Bradford connected dataset which is the main dataset in this work. The works describing the use of Bradford dataset were presented in Chapter 6 and 7. The access to the dataset was gained by working along side the research team that work on the same dataset. The process starts by series of discussion with the team to help with data understanding and the connectivity issue concerning the location of data access. The issue was to decide whether the data could be accessed from a university networked computer or from the Bradford Institute for Health Research (BIHR) computer. However, as the process of transferring the data to university in safely and secured manner may take longer than this work and complicated, the Bradford dataset was able to access via computer network in BIHR. It took several weeks after discussion with the team to acquire the access through an application called Research Passport. However, the duration of access to the dataset was limited to the three years and require extending the access due to unforeseen situation such as the Covid-19 pandemic crisis. The Research Passport for the Bradford data access was presented in Appendix A.2.

8.1.2 Data Analysis Environment

A stable data environment is the goal in any data analytics projects or work. It describes as the location used for managing, executing, collecting, and disseminating the work. This work was conducted in two different environments due to the confidential and privacy issue of the datasets. The work with MIMIC-III dataset was conducted using the University of Leeds computer provided for Postgraduate student in the School of Computing. It was created in the PostgreSQL Server in Windows 10 environment with 16.0 GB computer memory capacity. While the Bradford dataset was managed in the SQL Server in Windows 7 environment with similar computer memory capacity. However, most of the analysis work including data transformation were programmed using Python in Anaconda platform of openly source distribution exploiting Spyder application and Jupyter Notebook modules.

The most challenging part when working in big data area is the insufficient storage for or during data analysis. This challenge was commonly solved by requesting the IT department team (from the University of Leeds and BIHR) to

increase the storage capacity to reduce data analysis work abruption. The increased in storage capacity has proved to accelerate the data analysis work.

8.1.3 Quality of the Data

The most intimidating issue in any data analytics work is the quality of the data. It is even critical when utilizing the healthcare data for analytical work as the purpose of gathering the patient record for healthcare services and research work is different. The main goal for EHR is to facilitate the communication between healthcare professionals in delivering the best possible service to the patient. However, as the quality of the data for research was distrusted due to the distinctive goal, a comprehensive data quality checking and strategies required as the initial step in the work.

The source of data quality identified at several levels starting from the collection of the data management until the storing of the data. As the source of data quality was compromised, it affects the quality of value attributes of component in the event logs generated for process mining analysis such as event value, timestamp, resources, or case. In general, missing data, imprecise data, incorrect data, and irrelevant data are four main data quality issue in event log (Bose, Mans and Van Der Aalst, 2013). Example of commonly found quality issues in event log are incorrect timestamps, overlapping timestamps, missing events, incorrect event ordering, and inaccurate resource details (Martin *et al.*, 2019; Wynn and Sadiq, 2019).

The quality of the data was examined based on the framework outlined in (Kahn *et al.*, 2016). The framework facilitates in mitigating, identifying, and reporting the data quality issues. It comprises of three categories of data quality: conformance, completeness, and plausibility. The main goal of data quality inspection was to determine the applicability of the data for process mining analysis. The data quality inspection for MIMIC-III dataset has been discussed in Section 4.1.5 and Bradford dataset in Section 5.5. The benefit of applying data quality inspection was that the dataset produced was in high-quality of data gathered for analysis. However, the drawbacks of quality inspection were that rather smaller amount of data and often less representative of the population gathered. The data quality issue has been addressed differently with two datasets used in this work. To ensure more representative of the dataset is achieved, discussion with the domain experts and database management team was done, while this is not happened with MIMIC-III dataset as the access to hospital data is impossible.

8.1.4 Complexity of the Data

The complexity of the healthcare data is inarguably a challenging task to deal with. The typical healthcare data includes multi-spectral, incomplete, heterogeneous, and inaccurate observation (Dinov, 2016). It derived from variety of sources following inconsistent sampling during the delivery of healthcare services. Apart from that, the distinctive characteristics of big data specifically *Volume* (referring to the huge amount of the data collected within the EHR), *Variety* (referring to the presence of the data from various sources in a variety of formats), *Velocity* (referring to the speed of which the data is generated) increase the challenge in working with healthcare data (Martin-Sanchez and Verspoor, 2014).

The comprehension of the complexity of healthcare data could facilitates in understanding of the data for analysis work. The first dataset, MIMIC-III was provided with the dataset documentation from the website related to the development of the database, system, and extraction processes. An additional understanding of the MIMIC-III was done through the published paper using the similar dataset for various objectives. However, this approach in understanding the dataset is restricted and may be noncomprehensive as no direct access to hospital data and their expert domain is possible. Apart from that, the time shifting strategy applied on MIMIC-III for privacy and confidential purpose has limiting the analysis work. It affects the analysis related to the bottleneck of processes and performance related analysis.

Whereas the understanding of Bradford dataset was done through data exploration and discussion with the domain experts. The absence of explanation documentation related to the dataset extraction, structure, data collection strategies and EHR management has led to the persistent question-and-answer sessions with the domain experts and the dataset management team. In addition to that, the ever-changing data over time such as the coding standards has complicated the understanding of this complex data. The transitioning from the clinical coding of Read Code CTV3 to SNOMED which has been takes place in stages since April 2018 was an example of non-standardise coding scheme (Bradley, Lawrence and Carder, 2018). However, this transitioning was not affected in this study as the study duration was until end of year 2017 as discussed in Section 5.4.2. Further discussion in understanding the data was achieved and beneficial to conduct the analyses.

8.1.5 Visualisation of Findings

Apart from the three characteristics of big data discussed in previous section, the latest characteristics has emerged due to the growth in research work resulted from the increased of data availability. The *Value* is indicating as benefit and relevance of the date towards achieving specific outcomes of research work (Sedig and Ola, 2014). One of the outcomes is to acquire interactive visual representation of finding from data analytical work to illustrate the interrelation between variables and attributes of the data. The commonly used and basic kind of visual representation are comprising of line chart, bar chart and boxplot with whisker chart.

However, the fundamental of visual representation as mentioned above may be insufficient especially when it involved the knowledge transitioning between two domain areas. A useful and relevant data representation was achieved when it could facilitate the communication of the finding of work to the domain experts. Despite that, determining the best visualisation is an arduous work. Hence, in this work several visual representations was chosen to deliver the findings using process models (in multiple types), dotted chart, and trace variant.

The main deliverables of any process mining work are process model which can be represented in various form such as BPMN model, Petri net, transition system, process tree or causal net. The process models are commonly used to describe the workflow of an individual activity in any business process extracted from the information system to another activity in sequence manner. Different process mining tools generate different representation of process model, for example Disco produces a directly followed of fuzzy mining model, while ProM tools can generate various process model of which some of them are BPMN, petri net and transition system. Despite the user-friendly interface that Disco tool offer, the process model produces were suffered from low semantic model. This is because the algorithm adopted in Disco will only perceive the significant activity in the process by eliminating the non-significant activity thus resulting in less logical sequence of process model when compared to the real process execution. Hence, in this work most of the process model used was consists of model generated from ProM tool such as process tree in Section 4.2 (experiment 1) and transition system in Section 7.4.2 (experiment 7). However, the process model produced in Disco were still usable in this work if only the number of activities is acceptable (in terms of not generating a complex and low semantic model) as illustrates in Section 6.4 (experiment 5).

The dotted chart is a wider view of data visualisation to show the distribution of the data across durations of time such as in years, months, weeks etc. The chart represented using the presence of different attributes across time (e.g., study duration) in a dot shape with different colour. However, it has a drawback when too many different attributes were involved in the chart as the chart only captured about twenty-three different attributes to show. If more attributes involved, some of the colours of the dot may overlap. Hence, to overcome this issue an events abstraction or filtration approach may require on attributes number exceeding twenty-three.

Meanwhile trace variants present the sequence of events in grouped according to the proportion of cases sorting in ascending or descending order. This kind of visual representation is beneficial in illustrating the commonly and less common order of sequence observed in a log. However, it has limitation as the representation was only limited to show the proportion of the sequence but not on the performance attributes of the trace variant, such as the total or average duration each trace variant had.

8.2 Principle Finding based on Research Questions Assessment

The aim of this thesis is to analyse the frail elderly pathways with respect to frailty progression using the eFI. Comprehension of the dynamic of frailty progression is useful in informing current and future care needs, allowing for more prompt planning of suitable treatments, service design and labour requirements. The progression of frailty and its association with deficits of concern were explored in both datasets respectively. Furthermore, the present cut-off points followed in the literature (Clegg's approach) was compared with the proposed cut-off points in this study was also presented. Three approaches were applied to model the frailty progression and determine the cut-off points to identify different frailty stages. The approach in modelling frailty progression adopting the process mining techniques to best capture the variability of progression at different stages of frailty.

Two primary research questions have been presented in Section 1.6. They were broken down into five other RQs to provide as a guideline in this study. Seven (7) experiments were designed to answer the seven (7) RQ of this study. The approach and principle finding of each experiment in fulfilling the RQs of this

study are summarised. Table 8.1 presents the summary of approaches taken to conduct the analyses with its aim and deliverables respectively.

Table 8. 1 The summary of the fulfilment of the research questions of the work

RQ	Approach	Aim & Deliverables
RQs: 1, 2, 4	• Chapter 4 (MIMIC-III) • Experiment 1 • Work covered the hospital admission workflow using process mining technique	• Preliminary work in investigating the suitability of applying process mining • Process tree as a representation of the workflow hospital admission among elderly
RQs: 1, 2, 3, 4	• Chapter 4 (MIMIC-III) • Experiment 2 • Analysis covered the frailty trajectories using diagnosis code within elderly of different frailty categories	• Examine the suitability of applying process mining in analysing variability of frailty trajectories utilising the eFI • Directly followed model to illustrates the frailty trajectories and trace variant
RQs: 1, 2, 3, 4, 5	• Chapter 6 (Bradford dataset) • Experiment 3 • Covered the analysis of frailty trajectories using the diagnosis related code	• Analyse the differences of significant frailty trajectories • Directly followed model to illustrates the frailty trajectories
RQs: 1, 2, 3, 4, 5	• Chapter 6 (Bradford dataset) • Experiment 4 • Exploiting the process cube-based analysis approach to robustly dissect the frailty stages	• To explore the frailty progression over time within frailty stages • Significant pattern of frailty trajectories in the form of trace variant
RQs: 1, 2, 3, 4, 5	• Chapter 6 (Bradford dataset) • Experiment 5 • Included the deficits fall, hypertension and polypharmacy in the investigation of frailty progression	• To examine the frailty progression and its association with deficits of concern • Process model (directly followed of fuzzy-mined model and transition system)
RQS: 5, 6, 7	• Chapter 7 (Confirmatory analysis) • Experiment 6 • Applying the discretization approach in determining the cut-off points	• To determine the cut-off points to identify frailty stages based on target features • The proposed cut-off points between frailty stages fit, mild, moderate, and severe
RQS: 5, 6, 7	• Chapter 7 (Confirmatory analysis) • Experiment 7 • Comparative analysis between two cut-off points, Clegg's approach, and proposed approach as confirmatory analysis	• To reveal the diverse pattern of frailty progression within frailty stages based on different cut-off points approaches • Transition system as part of comparative analysis

8.2.1 RQ #1: Is it possible to analyse frail elderly pathway (frailty progression) using data and process mining analysis?

The possibility of applying process mining technique utilising the eFI score to analyse frailty pathway in the aspect of frailty progression was explored in the first five experiments of this thesis. The research question was answered by the preliminary work through experiment 1 in Section 4.2, examining the frailty trajectories using MIMIC-III dataset in experiment 2, Section 4.3, using Bradford

dataset in experiment 3 Section 6.2, exploring the frailty progression within frailty stages in experiment 4 Section 6.3, and its association with deficits of concern in experiment 5 Section 6.4.

8.2.2 RQ #2: Are the cut-off points used in eFI literature confirmed by real life data?

The work of confirmatory analysis through comparative analysis was presented in experiment 7 Section 7.4.2. The summary of the comparison point between two wors are presented in Table 8.2. The finding reveals that, there is slight different of the proposed cut off points of this work compared to the cut off points reported in the literature (Clegg *et al.*, 2016). The main reason to the difference found that mainly because of the aim of the two works. The work reported from the literature aims at developing the tool for identification of frailty severity. While in this work focusing in determining the optimal cut off points based on the target variables. The finding of optimal cut off points using a supervised discretization approach following a machine learning approach.

Table 8. 2 The comparison points between the approach followed in reported study and this work

Cut Off Points	Clegg & et al.	Proposed (this work)
Approach to Cut Off	Quantile (with 99% as upper limit)	Discretization
Distribution/Proportion (Prevalence Estimates)	43, 37, 16, 4 (based development cohort)	It portrays as binomial distribution although the literature reported left-skewed patient distribution.
Based On (Reference Value in Determining the Cut Off)	None (based on selection criteria of patient aged 65 years and over from the EHR)	Target variables (age, duration, events, contact)
Focus/Aim/Goal	Development of tools	Cut off points

The splitting is based on the maximum measure of divergence between the bins. On the note of Clegg et al. cut off points, the proportion of patient is left-skewed distribution. It describes as the highest proportion of elderly is in not frail category or fit, followed by mild, moderate and the least is in severe. However, in the proposed cut off point, the proportion of elderly shown in Figure 7.3 represented as binomial distribution, with the highest elderly is in moderate frail category. The justification to the difference is because the discretization used filtered dataset,

where the selection criteria with patient who has at least a year record. It is measured by the first and last unique deficits identified within the study duration. The approach had filtered out 65% (15,758) of patient from the dataset

8.2.3 RQ #3: What is the best illustration to portray frailty pathway?

The best illustration to present the finding was a challenge in this study. The decision to determine the best visual representation to deliver the finding of the work depends on the aim and viewpoint of the experiment. The best visual representation was decided after exploring all the possible and relevant representation each tools offered as discussed in Chapter 3. All experiments in this thesis have explored the possible visual representation to deliver the finding of the study. The visual representations are dotted chart, trace variant diagram, process models which includes transition system for log comparison.

8.2.4 RQ #4: Is it possible to analyse frailty progression utilizing electronic frailty index (eFI) scores?

The categorisation of frailty stages determined based on the eFI score has made the analysis of frailty progression possible. This has been answered using MIMIC-III dataset in Section 4.3 and using Bradford dataset in Section 6.2. The severity of frailty can be characterised as a workflow of transition into four different frailty stages: Fit (not frail), Mild, Moderate and Severe. The flow of transition is based to the level of severity from low to high. The frailty transition was analysed using time-based analysis employing the process mining approach. The progression of frailty is varying especially within different age range cohort as explored in experiment 4 in Section 6.3, and the initial frailty stages is having the longest time to move into the next stage as discussed in experiment 3 Section 6.2.

8.2.5 RQ #5: How can the dataset in relation to eFI score be extracted from the EHR system?

The extraction of the dataset was initially done within the EHR of elderly patient before identifying the events associated with the frailty, or frailty deficits as discussed in Section 5.8.1.2. The events were extracted and transformed through mapping the events into clinical classes and determining the frailty score when each deficit recorded within the EHR. Later, the next transformation done by completing the minimum requirement of an event log for process mining analysis.

It consists of case ID, activity, timestamps, and resources (frailty stage of the current location of the deficits). All these steps were demonstrated in Stage II (Extraction) and III (Data transformation and loading) of study method.

8.2.6 RQ #6: Is it possible to determine the new cut-off points following data-based approach?

The approach in identifying the cut-off points in determining the frailty stages was possible and demonstrated in experiment 6 Section 7.3. The discretization approach implementing the optimal binning method applied in this experiment has recognised four cut-off points based on the features found in the dataset. The experiment also found that proportion of frail elderly at each frailty categories different compared to the Clegg's approach.

8.2.7 RQ #7: What features can be used to characterise the new cut-off points in eFI score?

The Bradford dataset consists of four features to be included as the target variables in determining the optimal cut-off points for frailty scores. An exploratory data analysis on the available features in the dataset was presented in the Section 7.2. The features selected based on its association with each feature. The association coefficient was measured using the Spearman's' rank correlation test as presented in Section 7.3.3.3.2.1. The test found that the correlation value of all features was above 0.3 which is considered as having small correlation. Thus, the features a comprises of age of elderly at the final deficit, duration between the first and last deficit, number of healthcare professional contact, number of event (redundant deficits) and number of unique deficits.

8.3 Contributions to the Knowledge

The thesis goal is to contribute to the community of Process-Oriented Data Science for Health (PODS4H). The contribution the study was discussed which includes: (1) case studies in frail elderly in analysing frailty progression, (2) selection strategy for elderly cohort creation for analysis, (3) approach for analysing frailty progression and (4) determining the cut-off points for frailty score.

8.3.1 Application of process mining into frail elderly population

The first contribution of this study was the application of process mining techniques into the frail elderly population in two different case studies. The first case studies are MIMIC-III and Bradford dataset. To the best of our knowledge, this is the first work implemented process mining techniques into the elderly cohort of MIMIC-III and Bradford dataset. The secondary source of healthcare dataset which came from EHR often are not structured according to specific process workflow as the health condition of an individual is unique. Hence, working with this kind of data required a comprehensive step to evaluate its quality and improve the structure of dataset before starting the analysis. In this study, the steps discussed were explored in the data quality inspection step and data transformation as part of the data processing.

8.3.2 Selection Strategy for Elderly Cohort Creation

The second contribution in this study is the selection strategy in creating the frail elderly cohort which fit for the frailty progression analysis. Fit for the analysis defined as a subject of the study that has criteria which will reduce the bias towards the result. A careful selection strategy is required as the EHR constitutes the record within the study duration in which an individual health condition is varied. Hence, eliminating the assumption that frailty may already developed and advancing in any elderly patient at the start of the study duration is necessary.

8.3.3 Codes Mapping

The third contribution of the study is the codes mapping following the ontological concept approach as discussed in Section 5.8.1. The mapping of the activity events (frailty deficits) was done as part of the event abstraction approach. It facilitates in creating a simplified and understandable process model to deliver the finding of the work to the healthcare domain.

8.3.4 Approach for Analysing Frailty progression

Two approaches of exploring the frailty progression were done using MIMIC-III dataset (in Section 4.3) and using Bradford dataset (in Section 6.2) are the fourth

contribution in this study. The first method explores frailty progression based on the final frailty categories and the second method based on the frailty stages. Although, both methods successfully resulted in frailty trajectories, the second method however produced more comprehensive result as the analysis investigated the progression at similar points. The similar points refer to the progression of frailty from point of fit stage until reaching the next frailty stages, mild, from mild to moderate and moderate to severe. While the frailty categories compared the progression between frail elderly according to their final accumulated deficits. This method found significant pattern of trajectories at each frailty stages and the differences between stages.

The association between the deficits of concern and frailty progression has been explored and covered in Section 7.4. The association has been analysed using the process-cube and variant analysis to determine the dominant pattern sequence. The dominant pattern sequence is identified by a transition point from and to frailty stages. The association between deficits of concern and frailty progression is varies following certain pattern of sequence.

Although the MIMIC-III did not have the full deficits (e.g., polypharmacy) it still provides a basis work in modelling the frailty trajectories using main dataset, Bradford dataset. The modelling of frailty progression following the statistical steps to determine only the significant trajectories of one deficit to another within different frailty categories. The representation of the trajectories has been illustrated using the process mining approach using variety if plug-in and algorithm for both datasets.

8.3.5 Determining alternative cut-off points for frailty scores

The fifth contribution in this study is the alternative method of determining the cut-off points in identifying the frailty stages using existing eFI scores. The approach implementing the discretization technique which considered the available features in the dataset. The features were used to facilitate the method (discretization) in determining the cut-off points. The motivation behind this is that the question of whether the existing cut-off points scores had much difference with the proposed cut-off points was very much interesting to explore.

8.4 Limitation of the Study

Despite the successful analysis in investigating the frailty progression and its association with deficits of concern, there are some limitations that should be aware in this study. The limitations are as follow:

- Most of the process mining technique in this study focus on the process discovery. Less step-in conformance checking was done to measure the similarity between the reality and the assumption of the clinicians based on the disease pathway within the frail elderly.

- The study considered all events (Read codes) associated with frailty deficits. However, this study did not consider an approach in selecting the most prevalence frailty deficits (Read codes). The prevalence frailty deficits defined as the significant codes associated with frailty that being used by clinicians at certain point in time.

- One of the cohort creation strategies build on the assumption that the maximum frailty deficit increment in a year is 3. This may need in depth investigation as the increment could be varied depending on several factors such as different population of the cohort, age of the elderly and gender.

- This study only covered the Read code Version 3 (CTV3) which has been deprecated by the UK health provider. The conversion to SNOMED CT was taken place on April 2018 by stages. Although the study duration of this work is until 2017 which did not consider any SNOMED CT code within the analysis, it is not clear how well the approaches developed will work with SNOMED CT code in analysing frailty progression.

8.5 Future Direction

The work in this study has spaces for improvement in the future using the following ways:

- Exploring other definition of encompassing polypharmacy deficit such as with total number of medications ten. This is an interesting research work

as the comparison between different definition of polypharmacy would be worth to examine.

- Investigate the applicability of applying other machine learning technique in determining the optimal cut-off points of frailty scores as well as exploring the weakness and strength of each approaches taken.

- Association of other features as the outcome of the work in determining the cut-off points is worth to explore. The outcome features could be the mortality data of the patient, admission to the hospital and admission into institutionalisation for comparison work.

8.6 Summary

This chapter summarised the principle finding based on the research questions assessment. It also summarised the challenges faced while working in this research, the work contribution to the body of knowledge, its limitation and future direction as an improvement of the work. The challenges on working with healthcare data for process mining are related to many factors, including the data access and ethics approval, data quality, data understanding. Despite those challenges, the method proposed in this research has been successfully applied to the two datasets to analyse frailty progression and its association with deficits of concern.

References

Van der Aalst, W. M. . *et al.* (2012) 'Process mining manifesto', *BPM 2011 Workshops Proceedings*, pp. 169–194. doi: 10.1016/j.is.2011.10.006.

van der Aalst, W. M. P. (2011) *Process Mining—Discovery, Conformance and Enhancement of Business Processes*. Springer, Berlin. doi: https://doi.org/10.1007/978-3-642-19345-3.

van der Aalst, W. M. P. (2013) 'Process Cubes: Slicing, Dicing, Rolling Up and Drilling Down Event Data for Process Mining', in *Asia-Pacific Conference on Business Process Management. Lecture Notes in Business Information Processing*, pp. 1–22. doi: 10.1007/978-3-319-02922-1_1.

van der Aalst, W. M. P. (2015) 'Extracting Event Data from Databases to Unleash Process Mining', pp. 105–128. doi: 10.1007/978-3-319-14430-6_8.

Van der Aalst, W. M. P. and Weijters, A. J. M. M. (2004) 'Process mining: A research agenda', *Computers in Industry*, 53(3), pp. 231–244. doi: 10.1016/j.compind.2003.10.001.

Van der Aalst, W. (2016) *Process mining: Data science in action*, *Springer-Verlag Berlin Heidelberg*. doi: 10.1007/978-3-662-49851-4.

Van Der Aalst, W., Weijters, T. and Maruster, L. (2004) 'Workflow mining: Discovering process models from event logs', *IEEE Transactions on Knowledge and Data Engineering*, 16(9), pp. 1128–1142. doi: 10.1109/TKDE.2004.47.

Abbasi, M. *et al.* (2019) 'Correction to: A cross-sectional study examining convergent validity of a frailty index based on electronic medical records in a Canadian primary care program', *BMC Geriatrics*. BMC Geriatrics, 19(1), pp. 1–8. doi: 10.1186/s12877-019-1144-9.

Abu Bakar, A. A.-Z. *et al.* (2021) 'Older Adults with Hypertension: Prevalence of Falls and Their Associated Factors', *International Journal of Environmental Research and Public Health*, 18(16), p. 8257. doi: 10.3390/ijerph18168257.

Adeniran, R. (2004) 'The United Kingdom and United States Health Care Systems: a Comparison', *Home Health Care Management & Practice*, 16(2), pp. 109–116. doi: 10.1177/1084822303258617.

Ambagtsheer, R. C. *et al.* (2019) 'Application of an electronic Frailty Index in Australian primary care: data quality and feasibility assessment', *Aging Clinical and Experimental Research*. Springer International Publishing, 31(5), pp. 653–660. doi: 10.1007/s40520-018-1023-9.

Aubert, C. E. *et al.* (2016) 'Polypharmacy and specific comorbidities in university primary care settings', *European Journal of Internal Medicine*, 35, pp. 35–42. doi: 10.1016/j.ejim.2016.05.022.

Ayyar, A. *et al.* (2010) 'The journey of care for the frail older person', *British Journal of Hospital Medicine*, 71(2), pp. 92–96. doi: 10.12968/hmed.2010.71.2.46487.

Bannert, M., Reimann, P. and Sonnenberg, C. (2014) 'Process mining techniques for analysing patterns and strategies in students' self-regulated learning', *Metacognition and Learning*, 9(2), pp. 161–185. doi: 10.1007/s11409-013-9107-6.

Bartosch, P., McGuigan, F. E. and Akesson, K. E. (2018) 'Progression of frailty and prevalence of osteoporosis in a community cohort of older women—a 10-year

longitudinal study', *Osteoporosis International Journal*. Osteoporosis International, 29(10), pp. 2191–2199. doi: 10.1007/s00198-018-4593-7.
Beam, A. L. and Kohane, I. S. (2018) 'Big Data and Machine Learning in Health Care', *JAMA*, 319(13), p. 1317. doi: 10.1001/jama.2017.18391.
Beard, J. R. and Bloom, D. E. (2015) 'Towards a comprehensive public health response to population ageing', *The Lancet*, 385(9968), pp. 658–661. doi: 10.1016/S0140-6736(14)61461-6.
Bezerra, A. *et al.* (2019) 'Extracting value from industrial alarms and events: A data-driven approach based on exploratory data analysis', *Sensors (Switzerland)*, 19(12). doi: 10.3390/s19122772.
BIDMC (2020) *About Beth Deaconess Medical Center*, *Online*. Available at: https://www.bidmc.org/about-bidmc (Accessed: 21 December 2020).
BMP (2019) *About Bingley Medical Practice*, *Online*. Available at: https://www.bingleymedical.org.uk/about-us/ (Accessed: 22 December 2020).
Bock, J.-O. *et al.* (2016) 'Associations of frailty with health care costs – results of the ESTHER cohort study', *BMC Health Services Research*, 16(1), p. 128. doi: 10.1186/s12913-016-1360-3.
Bolt, A., de Leoni, M. and van der Aalst, W. M. P. (2018) 'Process variant comparison: Using event logs to detect differences in behavior and business rules', *Information Systems*. Elsevier Ltd, 74, pp. 53–66. doi: 10.1016/j.is.2017.12.006.
Bolt, A., De Leoni, M. and Van Der Aalst, W. M. P. (2016) 'A visual approach to spot statistically-significant differences in event logs based on process metrics', *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 9694, pp. 151–166. doi: 10.1007/978-3-319-39696-5.10.
Bose, R. P. J. C. and Aalst, W. M. P. Van Der (2013) 'Dealing With Concept Drifts in Process Mining', pp. 1–18.
Bose, R. P. J. C., Aalst, W. M. P. Van Der and Pechenizkiy, M. (2011) 'Handling Concept Drift in Process Mining', pp. 391–405. Available at: http://65.54.113.26/Publication/48806852/handling-concept-drift-in-process-mining.
Bose, R. P. J. C., Mans, R. S. and Van Der Aalst, W. M. P. (2013) '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*, (April), pp. 127–134. doi: 10.1109/CIDM.2013.6597227.
Bradley, S. H., Lawrence, N. R. and Carder, P. (2018) 'Using primary care data for health research in England – an overview', *Future Healthcare Journal*, 5(3), pp. 207–212. doi: 10.7861/futurehosp.5-3-207.
Broad, A. *et al.* (2020) 'The convergent validity of the electronic frailty index (EFI) with the clinical frailty scale (CFS)', *Geriatrics (Switzerland)*, 5(4), pp. 1–6. doi: 10.3390/geriatrics5040088.
Bromfield, S. G. *et al.* (2017) 'Blood Pressure, Antihypertensive Polypharmacy, Frailty, and Risk for Serious Fall Injuries Among Older Treated Adults With Hypertension', *Hypertension*, 70(2), pp. 259–266. doi: 10.1161/HYPERTENSIONAHA.116.09390.
Brundle, C. *et al.* (2018) 'Convergent validity of the electronic frailty index.', *Age and ageing*, pp. 1–5. doi: 10.1093/ageing/afy162.

Buja, A. *et al.* (2020) 'Healthcare Service Usage and Costs for Elderly Patients with Obstructive Lung Disease', *International Journal of Chronic Obstructive Pulmonary Disease*, Volume 15, pp. 3357–3366. doi: 10.2147/COPD.S275687.
Cai, A. and Calhoun, D. A. (2018) 'Antihypertensive Medications and Falls in the Elderly', *American Journal of Hypertension*, 31(3), pp. 281–283. doi: 10.1093/ajh/hpx203.
Callahan, K. E. *et al.* (2021) 'Automated Frailty Screening At-Scale for Pre-Operative Risk Stratification Using the Electronic Frailty Index', *Journal of the American Geriatrics Society*, 69(5), pp. 1357–1362. doi: 10.1111/jgs.17027.
Callisaya, M. L. *et al.* (2014) 'Greater Daily Defined Dose of Antihypertensive Medication Increases the Risk of Falls in Older People-A Population-Based Study', *Journal of the American Geriatrics Society*, 62(8), pp. 1527–1533. doi: 10.1111/jgs.12925.
Chamberlain, A. M. *et al.* (2016) 'Frailty trajectories in an elderly population-based cohort', *Journal of the American Geriatrics Society*, 64(2), pp. 285–292. doi: 10.1111/jgs.13944.
Cheng, M.-H. and Chang, S.-F. (2017) 'Frailty as a Risk Factor for Falls Among Community Dwelling People: Evidence From a Meta-Analysis', *Journal of Nursing Scholarship*, 49(5), pp. 529–536. doi: 10.1111/jnu.12322.
Christian, W. G. *et al.* (2008) 'Using Process Mining to Learn from Process Changes in Evolutionary Systems'.
Clegg, A. *et al.* (2013) 'Frailty in elderly people', *The Lancet*, 381(9868), pp. 752–762. doi: 10.1016/S0140-6736(12)62167-9.
Clegg, A. *et al.* (2016) 'Development and validation of an electronic frailty index using routine primary care electronic health record data', *Journal of Age and Ageing*, 45(3), pp. 353–360. doi: 10.1093/ageing/afw039.
Conca, T., Saint-Pierre, C., Herskovic, V., Sepúlveda, M., Capurro, D., Prieto, F. and Fernandez-Llatas, C., 2018. Multidisciplinary Collaboration in the Treatment of Patients with Type 2 Diabetes in Primary Care: Analysis Using Process Mining. Journal of medical Internet research, 20(4).
Cowie, M. R. *et al.* (2017) 'Electronic health records to facilitate clinical research', *Clinical Research in Cardiology*, 106(1), pp. 1–9. doi: 10.1007/s00392-016-1025-6.
Crimmins, E. M. and Beltran-Sanchez, H. (2011) 'Mortality and Morbidity Trends: Is There Compression of Morbidity?', *The Journals of Gerontology Series B: Psychological Sciences and Social Sciences*, 66B(1), pp. 75–86. doi: 10.1093/geronb/gbq088.
Cuadrado, M. T. (2019) *ICD-9-CM: International Classification of Diseases, Ninth Revision, Clinical Modification*, *Online*. Available at: https://ec.europa.eu/cefdigital/wiki/display/EHSEMANTIC/ICD-9-CM%3A+International+Classification+of+Diseases%2C+Ninth+Revision%2C+Clinical+Modification (Accessed: 21 December 2020).
Curcio, C.-L., Henao, G.-M. and Gomez, F. (2014) 'Frailty among rural elderly adults', *BMC Geriatrics*, 14(1), p. 2. doi: 10.1186/1471-2318-14-2.
Dagli, R. J. and Sharma, A. (2014) 'Polypharmacy : A Global Risk Factor for Elderly People', 6(6), pp. 6–7.
Dash, R., Paramguru, R. L. and Dash, R. (2011) 'Comparative analysis of supervised and unsupervised discretization techniques', *nternational Journal of Advances in Science and Technology*, 2(3).
Dash, Rajashree, Paramguru, R. L. and Dash, Rasmita (2011) 'Comparative Analysis of

Supervised and Unsupervised Discretization Techniques', *International Journal of Advances in Science and Technology*, 2(3).
Delespierre, T. *et al.* (2017) 'Empirical advances with text mining of electronic health records', *BMC Medical Informatics and Decision Making*, 17(1), p. 127. doi: 10.1186/s12911-017-0519-0.
Dent, E., Kowal, P. and Hoogendijk, E. O. (2016) 'Frailty measurement in research and clinical practice: A review', *European Journal of Internal Medicine*. European Federation of Internal Medicine, 31, pp. 3–10. doi: 10.1016/j.ejim.2016.03.007.
Dinov, I. D. (2016) 'Volume and value of big healthcare data', *Journal of Medical Statistics and Informatics*, 4(1), p. 3. doi: 10.7243/2053-7662-4-3.
Easton, J. F., Stephens, C. R. and Angelova, M. (2014) 'Risk factors and prediction of very short term versus short/intermediate term post-stroke mortality: A data mining approach', *Computers in Biology and Medicine*, 54, pp. 199–210. doi: 10.1016/j.compbiomed.2014.09.003.
Van Eck, M. L. *et al.* (2015) 'PM2: A process mining project methodology', *International Conference on Advanced Information Systems Engineering*, 9097, pp. 297–313. doi: 10.1007/978-3-319-19069-3_19.
Eeles, E. M. P. *et al.* (2012) 'The impact of frailty and delirium on mortality in older inpatients', *Age and Ageing*, 41(3), pp. 412–416. doi: 10.1093/ageing/afs021.
Eendebak, R. and Organization, W. H. (2015) 'World Report on Ageing and Health', *World Health Organization, Global*, p. 260. Available at: https://www.who.int/ageing/events/world-report-2015-launch/en/.
Espinoza, S. E., Jung, I. and Hazuda, H. (2012) 'Frailty Transitions in the San Antonio Longitudinal Study of Aging', *Journal of the American Geriatrics Society*, 60(4), pp. 652–660. doi: 10.1111/j.1532-5415.2011.03882.x.
Farid, N. F., De Kamps, M. and Johnson, O. A. (2019) 'Process mining in frail elderly care: A literature review', *HEALTHINF 2019 - 12th International Conference on Health Informatics, Proceedings; Part of 12th International Joint Conference on Biomedical Engineering Systems and Technologies, BIOSTEC 2019*, pp. 332–339.
Fhon, J. R. S. *et al.* (2016) 'Fall and its association with the frailty syndrome in the elderly: systematic review with meta-analysis', *Revista da Escola de Enfermagem da USP*, 50(6), pp. 1005–1013. doi: 10.1590/s0080-623420160000700018.
Finkel, D., Whitfield, K. and McGue, M. (1995) 'Genetic and Environmental Influences on Functional Age: A Twin Study', *The Journals of Gerontology Series B: Psychological Sciences and Social Sciences*, 50B(2), pp. P104–P113. doi: 10.1093/geronb/50B.2.P104.
Flegel, K. (2015) 'Tertiary hospitals must provide general care', *Canadian Medical Association Journal*, 187(4), pp. 235–235. doi: 10.1503/cmaj.150056.
Fogg, C. *et al.* (2022) 'The dynamics of frailty development and progression in older adults in primary care in England (2006–2017): a retrospective cohort profile', *BMC Geriatrics*. BioMed Central, 22(1), pp. 1–11. doi: 10.1186/s12877-021-02684-y.
Fried, L. P. *et al.* (2001) 'Frailty in Older Adults: Evidence for a Phenotype', *The Journals of Gerontology Series A: Biological Sciences and Medical Sciences*, 56(3), pp. M146–M157. doi: 10.1093/gerona/56.3.M146.
Fried, L. P. *et al.* (2004) 'Untangling the Concepts of Disability, Frailty, and Comorbidity: Implications for Improved Targeting and Care', *The Journals of Gerontology Series A:*

Biological Sciences and Medical Sciences, 59(3), pp. M255–M263. doi: 10.1093/gerona/59.3.M255.
Gielen, E. *et al.* (2012) 'Musculoskeletal Frailty: A Geriatric Syndrome at the Core of Fracture Occurrence in Older Age', *Calcified Tissue International*, 91(3), pp. 161–177. doi: 10.1007/s00223-012-9622-5.
Gill, T. M. *et al.* (2006) 'Transitions between frailty states among community-living older persons.', *Arch Intern Med*, 166(4), pp. 418–423. doi: 10.1001/archinte.166.4.418.
Gupta, A., Mehrotra, K. G. and Mohan, C. (2010) 'A clustering-based discretization for supervised learning', *Statistics & Probability Letters*, 80(9–10), pp. 816–824. doi: 10.1016/j.spl.2010.01.015.
Haux, R. (2006) 'Health information systems – past, present, future', *International Journal of Medical Informatics*, 75(3–4), pp. 268–281. doi: 10.1016/j.ijmedinf.2005.08.002.
Hesselink, G. *et al.* (2012) 'Improving Patient Handovers From Hospital to Primary Care', *Annals of Internal Medicine*, 157(6), p. 417. doi: 10.7326/0003-4819-157-6-201209180-00006.
Hogan, D. B. (2018) 'Models, Definitions, and Criteria for Frailty', in *Conn's Handbook of Models for Human Aging*. Elsevier, pp. 35–44. doi: 10.1016/B978-0-12-811353-0.00003-8.
Hoogendijk, E. O. *et al.* (2019) 'Frailty: implications for clinical practice and public health', *The Lancet*, 394(10206), pp. 1365–1375. doi: 10.1016/S0140-6736(19)31786-6.
Hoogendijk, E. O. *et al.* (2020) 'Operationalization of a frailty index among older adults in the InCHIANTI study: predictive ability for all-cause and cardiovascular disease mortality', *Aging Clinical and Experimental Research*. Springer International Publishing, 32(6), pp. 1025–1034. doi: 10.1007/s40520-020-01478-3.
Hsu, H.-C. and Chang, W.-C. (2015) 'Trajectories of Frailty and Related Factors of the Older People in Taiwan', *Experimental Aging Research*, 41(1), pp. 104–114. doi: 10.1080/0361073X.2015.978219.
Hubbard, R. E., O'Mahony, M. S. and Woodhouse, K. W. (2008) 'Characterising frailty in the clinical setting--a comparison of different approaches', *Age and Ageing*, 38(1), pp. 115–119. doi: 10.1093/ageing/afn252.
Ilinca, S. and Calciolari, S. (2015) 'The Patterns of Health Care Utilization by Elderly Europeans: Frailty and Its Implications for Health Systems', *Health Services Research*, 50(1), pp. 305–320. doi: 10.1111/1475-6773.12211.
Jagger, C. *et al.* (2008) 'Inequalities in healthy life years in the 25 countries of the European Union in 2005: a cross-national meta-regression analysis', *The Lancet*, 372(9656), pp. 2124–2131. doi: 10.1016/S0140-6736(08)61594-9.
Jans, M. *et al.* (2011) 'A business process mining application for internal transaction fraud mitigation', *Expert Systems with Applications*, 38(10), pp. 13351–13359. doi: 10.1016/j.eswa.2011.04.159.
Jans, M., Alles, M. G. and Vasarhelyi, M. A. (2014) 'A Field Study on the Use of Process Mining of Event Logs as an Analytical Procedure in Auditing', *The Accounting Review*, 89(5), pp. 1751–1773. doi: 10.2308/accr-50807.
Jebb, A. T., Parrigon, S. and Woo, S. E. (2017) 'Exploratory data analysis as a foundation of inductive research', *Human Resource Management Review*. Elsevier Inc., 27(2), pp. 265–276. doi: 10.1016/j.hrmr.2016.08.003.

Jensen, A. B. *et al.* (2014) 'Temporal disease trajectories condensed from population-wide registry data covering 6.2 million patients', *Nature Communications*, 5(May), pp. 1–10. doi: 10.1038/ncomms5022.
Johnson, A. E. W. *et al.* (2016) 'MIMIC-III, a freely accessible critical care database', *Scientific Data*, 3, pp. 1–9. doi: 10.1038/sdata.2016.35.
Jolliffe, I. T. and Cadima, J. (2016) 'Principal component analysis: a review and recent developments', *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 374(2065), p. 20150202. doi: 10.1098/rsta.2015.0202.
Jonas, M., Kazarski, R. and Chernin, G. (2018) 'Ambulatory blood-pressure monitoring, antihypertensive therapy and the risk of fall injuries in elderly hypertensive patients.', *Journal of geriatric cardiology : JGC*, 15(4), pp. 284–289. doi: 10.11909/j.issn.1671-5411.2018.04.007.
Jordan, M. I. and Mitchell, T. M. (2015) 'Machine learning: Trends, perspectives, and prospects', *Science*, 349(6245), pp. 255–260. doi: 10.1126/science.aaa8415.
Kahn, M. G. *et al.* (2016) 'A Harmonized Data Quality Assessment Terminology and Framework for the Secondary Use of Electronic Health Record Data', *eGEMs (Generating Evidence & Methods to improve patient outcomes)*, 4(1), p. 18. doi: 10.13063/2327-9214.1244.
Kidd, T. *et al.* (2019) 'What are the most effective interventions to improve physical performance in pre-frail and frail adults? A systematic review of randomised control trials', *BMC Geriatrics*, 19(1), p. 184. doi: 10.1186/s12877-019-1196-x.
Kim, E. *et al.* (2019) 'The Evolving Use of Electronic Health Records (EHR) for Research', *Seminars in Radiation Oncology*, 29(4), pp. 354–361. doi: 10.1016/j.semradonc.2019.05.010.
Kojima, G. *et al.* (2015) 'Frailty predicts short-term incidence of future falls among British community-dwelling older people: A prospective cohort study nested within a randomised controlled trial Physical functioning, physical health and activity', *BMC Geriatrics*. BMC Geriatrics, 15(1), pp. 1–8. doi: 10.1186/s12877-015-0152-7.
Kojima, G., Liljas, A. and Iliffe, S. (2019) 'Frailty syndrome: implications and challenges for health care policy', *Risk Management and Healthcare Policy*, Volume 12, pp. 23–30. doi: 10.2147/RMHP.S168750.
Kojima, T. *et al.* (2011) 'Association of polypharmacy with fall risk among geriatric outpatients', *Geriatrics & Gerontology International*, 11(4), pp. 438–444. doi: 10.1111/j.1447-0594.2011.00703.x.
Kurniati, A. P. *et al.* (2019) 'The assessment of data quality issues for process mining in healthcare using Medical Information Mart for Intensive Care III, a freely available e-health record database', *Health Informatics Journal*, 25(4), pp. 1878–1893. doi: 10.1177/1460458218810760.
Kusuma, G. P. *et al.* (2020) 'Process mining of disease trajectories: A feasibility study', in *HEALTHINF 2020 - 13th International Conference on Health Informatics, Proceedings; Part of 13th International Joint Conference on Biomedical Engineering Systems and Technologies, BIOSTEC 2020*, pp. 705–712. doi: 10.5220/0009166607050712.
Kusuma, G. P. *et al.* (2021) 'Process mining of disease trajectories: A literature review', *Public Health and Informatics: Proceedings of MIE 2021*, 0, pp. 457–461. doi: 10.3233/SHTI210200.

Lally, F. and Crome, P. (2007) 'Understanding frailty', *Postgraduate Medical Journal*, 83(975), pp. 16–20. doi: 10.1136/pgmj.2006.048587.
Lang, P. O., Michel, J. P. and Zekry, D. (2009) 'Frailty syndrome: A transitional state in a dynamic process', *Journal of Gerontology*, 55(5), pp. 539–549. doi: 10.1159/000211949.
Lansbury, L. N. *et al.* (2017) 'Use of the electronic Frailty Index to identify vulnerable patients: A pilot study in primary care', *British Journal of General Practice*, 67(664), pp. e751–e756. doi: 10.3399/bjgp17X693089.
Lee, S. B. *et al.* (2018) 'Differences in youngest-old, middle-old, and oldest-old patients who visit the emergency department.', *Clinical and experimental emergency medicine*, 5(4), pp. 249–255. doi: 10.15441/ceem.17.261.
Leemans, S. J. J., Poppe, E. and Wynn, M. T. (2019) 'Directly follows-based process mining: A tool', *CEUR Workshop Proceedings*, 2374, pp. 9–12.
Liu, H. *et al.* (2002) 'Discretization: An Enabling Technique', *Data Mining and Knowledge Discovery*, 6, pp. 393–423. doi: https://doi.org/10.1023/A:1016304305535.
Liu, P. *et al.* (2020) 'Frailty and hypertension in older adults: current understanding and future perspectives', *Hypertension Research*, 43(12), pp. 1352–1360. doi: 10.1038/s41440-020-0510-5.
Mannhardt, F., De Leoni, M. and Reijers, H. A. (2017) 'Heuristic mining revamped: An interactive, data-Aware, and conformance-Aware miner', *CEUR Workshop Proceedings*, 1920(August).
Mans, R. S. *et al.* (2013) 'Process Mining in Healthcare: Data Challenges When Answering Frequently Posed Questions', in *Process Support and Knowledge Representation in Health Care. ProHealth 2012, KR4HC 2012. Lecture Notes in Computer Science*. Springer (Lecture Notes in Computer Science), pp. 140–153. doi: 10.1007/978-3-642-36438-9_10.
Manton, K. G., Gu, X. and Lamb, V. L. (2006) 'Change in chronic disability from 1982 to 2004/2005 as measured by long-term changes in function and health in the U.S. elderly population', *Proceedings of the National Academy of Sciences*, 103(48), pp. 18374–18379. doi: 10.1073/pnas.0608483103.
Martin-Sanchez, F. and Verspoor, K. (2014) 'Big Data in Medicine Is Driving Big Changes', *Yearbook of Medical Informatics*, 23(01), pp. 14–20. doi: 10.15265/IY-2014-0020.
Martin, N. *et al.* (2019) 'Interactive Data Cleaning for Process Mining: A Case Study of an Outpatient Clinic's Appointment System', *Lecture Notes in Business Information Processing*, 362 LNBIP, pp. 532–544. doi: 10.1007/978-3-030-37453-2_43.
Masnoon, N. *et al.* (2017) 'What is polypharmacy? A systematic review of definitions', *BMC Geriatrics*. BMC Geriatrics, 17(1), pp. 1–10. doi: 10.1186/s12877-017-0621-2.
Menachemi, N. and Collum, T. H. (2011) 'Benefits and drawbacks of electronic health record systems.', *Risk management and healthcare policy*, 4, pp. 47–55. doi: 10.2147/RMHP.S12985.
Mendiratta, P. and Latif, R. (2021) *Clinical Frailty Scale*, *StatPearls*. Available at: http://www.ncbi.nlm.nih.gov/pubmed/32644435.
Morgenthaler, S. (2009) 'Exploratory data analysis', *Wiley Interdisciplinary Reviews: Computational Statistics*, 1(1), pp. 33–44. doi: 10.1002/wics.2.
Morris, J. A. and Gardner, M. J. (1988) 'Statistics in Medicine: Calculating confidence

intervals for relative risks (odds ratios) and standardised ratios and rates', *BMJ*, 296(6632), pp. 1313–1316. doi: 10.1136/bmj.296.6632.1313.
Mortazavi, S. S. *et al.* (2016) 'Defining polypharmacy in the elderly: A systematic review protocol', *BMJ Open*, 6(3), pp. 1–4. doi: 10.1136/bmjopen-2015-010989.
Mukaka, M. M. (2012) 'Statistics corner: A guide to appropriate use of correlation coefficient in medical research.', *Malawi medical journal : the journal of Medical Association of Malawi*, 24(3), pp. 69–71. Available at: http://www.ncbi.nlm.nih.gov/pubmed/23638278.
Munstermann, M., Stevens, T. and Luther, W., 2012. A Novel Human Autonomy Assessment System. Sensors, 12(6), pp.7828-7854.
Najjar, A., Reinharz, D., Girouard, C. and Gagné, C., 2018. A Two-step Approach for Mining Patient Treatment Pathways in Administrative Healthcare Databases. Artificial Intelligence in Medicine, 87, pp.34-48.
Nanayakkara, S. *et al.* (2014) 'Admission time to hospital: a varying standard for a critical definition for admissions to an intensive care unit from the emergency department', *Australian Health Review*, 38(5), p. 575. doi: 10.1071/AH13244.
Navas-Palencia, G. (2020) 'Optimal binning: mathematical programming formulation', *Computer Science, Mathematics*. Available at: http://arxiv.org/abs/2001.08025.
Nicholson, A. and Stone, B. (2013) 'Evidence to the Royal Commission on the National Health Service. 1. From the Clinical Pharmacology Section of the British Pharmacological Society.', *British Journal of Clinical Pharmacology*, 5(6), pp. 475–480. doi: 10.1111/j.1365-2125.1978.tb01659.x.
Nobili, A. *et al.* (2011) 'Association between clusters of diseases and polypharmacy in hospitalized elderly patients: Results from the REPOSI study', *European Journal of Internal Medicine*. European Federation of Internal Medicine, 22(6), pp. 597–602. doi: 10.1016/j.ejim.2011.08.029.
Nwadiugwu, M. C. (2020) 'Frailty and the Risk of Polypharmacy in the Older Person: Enabling and Preventative Approaches', *Journal of Aging Research*, 2020, pp. 1–6. doi: 10.1155/2020/6759521.
Oliver, D. (2009) 'Age based discrimination in health and social care services', *BMJ (Online)*, 339(7722), p. 643. doi: 10.1136/bmj.b3378.
Orfanidis, L., Bamidis, P. D. and Eaglestone, B. (2004) 'Data quality issues in electronic health records: an adaptation framework for the Greek health system', *Health Informatics Journal*, 10(1), pp. 23–36. doi: 10.1177/146045804040665.
Palangkaraya, A. and Yong, J. (2009) 'Population ageing and its implications on aggregate health care demand: empirical evidence from 22 OECD countries', *International Journal of Health Care Finance and Economics*, 9(4), pp. 391–402. doi: 10.1007/s10754-009-9057-3.
Palaniappan, S. and Hong, T. K. (2008) 'Discretization of continuous valued dimensions in OLAP data cubes', *International Journal of Computer Science and Network Security*, 8(1), pp. 116–126.
Pandey, S. (2018) 'Factors Contributing of Ageing', in *Handbook of Research on Geriatric Health, Treatment, and Care*, pp. 393–408. doi: 10.4018/978-1-5225-3480-8.ch022.
Pedregosa, F. *et al.* (2011) 'Scikit-learn: Machine Learning in Python', *JMLR*, 12(85), pp. 2825–2830.
Pialoux, T., Goyard, J. and Lesourd, B. (2012) 'Screening tools for frailty in primary health

care: A systematic review', *Journal of Geriatrics & Gerontology*, 12(2), pp. 189–197. doi: 10.1111/j.1447-0594.2011.00797.x.
Primary, I. *et al.* (2005) 'Codes , classifications , terminologies and nomenclatures : definition , development and application in practice A theme of the European Federation for Medical The Primary Care Informatics Working Group of', *Informatics in Primary Care*, pp. 65–69.
Puts, M. T. E. *et al.* (2017) 'Interventions to prevent or reduce the level of frailty in community-dwelling older adults: a scoping review of the literature and international policies.', *Age and ageing*, 46(3), pp. 383–392. doi: 10.1093/ageing/afw247.
Rieckert, A. *et al.* (2018) 'Polypharmacy in older patients with chronic diseases: A cross-sectional analysis of factors associated with excessive polypharmacy', *BMC Family Practice*. BMC Family Practice, 19(1), pp. 1–9. doi: 10.1186/s12875-018-0795-5.
Ritt, M. *et al.* (2015) 'Analysis of Rockwood et Al's Clinical Frailty Scale and Fried et Al's Frailty Phenotype as Predictors of Mortality and Other Clinical Outcomes in Older Patients Who Were Admitted to a Geriatric Ward', *The journal of nutrition, health & aging*, 19(10), pp. 1043–1048. doi: 10.1007/s12603-015-0534-8.
Rizzuto, D. *et al.* (2012) 'Lifestyle, social factors, and survival after age 75: population based study', *BMJ*, 345(aug29 2), pp. e5568–e5568. doi: 10.1136/bmj.e5568.
Rockwood, K. (2005) 'Frailty and Its Definition: A Worthy Challenge', *Journal of the American Geriatrics Society*, 53(6), pp. 1069–1070. doi: 10.1111/j.1532-5415.2005.53312.x.
Rockwood, K., Andrew, M. and Mitnitski, A. (2007) 'A comparison of two approaches to measuring frailty in elderly people', *Journals of Gerontology - Series A Biological Sciences and Medical Sciences*, 62(7), pp. 738–743. doi: 10.1093/gerona/62.7.738.
Rockwood, K. and Mitnitski, A. (2011) 'Frailty Defined by Deficit Accumulation and Geriatric Medicine Defined by Frailty', *Clinics in Geriatric Medicine*. Elsevier Ltd, 27(1), pp. 17–26. doi: 10.1016/j.cger.2010.08.008.
Rojas, E., Arias, M. and Sepúlveda, M. (2015) 'Clinical Processes and Its Data, What Can We Do with Them?', in *Proceedings of the International Conference on Health Informatics*. SCITEPRESS - Science and and Technology Publications, pp. 642–647. doi: 10.5220/0005287206420647.
Romero-Ortuno, R. (2013) 'An alternative method for Frailty Index cut-off points to define frailty categories', *European Geriatric Medicine*, 4(5), pp. 299–303. doi: 10.1016/j.eurger.2013.06.005.
Romero-Ortuno, R. and Kenny, R. A. (2012) 'The frailty index in Europeans: association with age and mortality', *Age and Ageing*, 41(5), pp. 684–689. doi: 10.1093/ageing/afs051.
Russler, D. (2009) 'Clinical Observation', in *Encyclopedia of Database Systems*. Boston, MA: Springer US, pp. 359–360. doi: 10.1007/978-0-387-39940-9_61.
Salinas-Rodríguez, A. *et al.* (2019) 'Healthcare Costs of Frailty: Implications for Long-term Care', *Journal of the American Medical Directors Association*, 20(1), pp. 102-103.e2. doi: 10.1016/j.jamda.2018.09.019.
Schober, P., Boer, C. and Schwarte, L. A. (2018) 'Correlation Coefficients: Appropriate Use and Interpretation', *Anesthesia & Analgesia*, 126(5), pp. 1763–1768. doi: 10.1213/ANE.0000000000002864.
Searle, S. D. *et al.* (2008) 'A standard procedure for creating a frailty index.', *BMC*

geriatrics, 8, p. 24. doi: 10.1186/1471-2318-8-24.
Sedig, K. and Ola, O. (2014) 'The Challenge of Big Data in Public Helth: An Opportunity for Visual Analytics', *Online Journal of Public Health Informatics*, 5(3). doi: 10.5210/ojphi.v5i3.4933.
Setiati, S. *et al.* (2019) 'Frailty state among Indonesian elderly: Prevalence, associated factors, and frailty state transition', *BMC Geriatrics*. BMC Geriatrics, 19(1), pp. 1–10. doi: 10.1186/s12877-019-1198-8.
Shalev-Shwartz, Shai, and S. B.-D. (2014) *Understanding machine learning: From theory to algorithms*. Cambridge University press.
Shim, E. Y. *et al.* (2011) 'Correlation between Frailty Level and Adverse Health-related Outcomes of Community-Dwelling Elderly, One Year Retrospective Study', *Korean Journal of Family Medicine*, 32(4), p. 249. doi: 10.4082/kjfm.2011.32.4.249.
Shrivastava, S. R. B. L., Shrivastava, P. S. and Ramasamy, J. (2013) 'Health-care of Elderly: Determinants, Needs and Services.', *International journal of preventive medicine*, 4(10), pp. 1224–5. Available at: http://www.ncbi.nlm.nih.gov/pubmed/24319566.
Sieber, C. C. (2017) 'Frailty – From concept to clinical practice', *Experimental Gerontology*. Elsevier B.V., 87, pp. 160–167. doi: 10.1016/j.exger.2016.05.004.
Simpson, K. N. *et al.* (2018) 'Effect of frailty on resource use and cost for Medicare patients', *Journal of Comparative Effectiveness Research*, 7(8), pp. 817–825. doi: 10.2217/cer-2018-0029.
Song, X., Mitnitski, A. and Rockwood, K. (2010) 'Prevalence and 10-Year Outcomes of Frailty in Older Adults in Relation to Deficit Accumulation', *Journal of the American Geriatrics Society*, 58(4), pp. 681–687. doi: 10.1111/j.1532-5415.2010.02764.x.
Stewart, S. T., Cutler, D. M. and Rosen, A. B. (2013) 'US Trends in Quality-Adjusted Life Expectancy From 1987 to 2008: Combining National Surveys to More Broadly Track the Health of the Nation', *American Journal of Public Health*, 103(11), pp. e78–e87. doi: 10.2105/AJPH.2013.301250.
Stöber, J. *et al.* (2015) 'Comorbidity of chronic diseases in the elderly: Patterns identified by a copula design for mixed responses', *Computational Statistics & Data Analysis*, 88, pp. 28–39. doi: 10.1016/j.csda.2015.02.001.
Stow, D. *et al.* (2018) 'Evaluating frailty scores to predict mortality in older adults using data from population based electronic health records: case control study', *Age and Ageing*, pp. 564–569. doi: 10.1093/ageing/afy022.
Stow, D., Matthews, F. E. and Hanratty, B. (2018) 'Frailty trajectories to identify end of life: A longitudinal population-based study', *BMC Medicine*. BMC Medicine, 16(1), pp. 1–7. doi: 10.1186/s12916-018-1148-x.
Strehl, V. (2013) 'Clinical consequences of polypharmacy in Elderly', 13(1), pp. 1–11. doi: 10.1517/14740338.2013.827660.Clinical.
Suzman, R., Beard, J. R. and Organization, W. H. (2011) 'Global Health and Aging', *World Health Organization, Global*, p. 32. Available at: https://www.who.int/ageing/publications/global_health/en/.
SWMP (2021) *About the Saltaire and Windhill Medical Partnership*, *Online*. Available at: https://saltairewindhillgp.org/about/primary-care-network/ (Accessed: 5 January 2021).
Tarekegn, A. *et al.* (2020) 'Predictive Modeling for Frailty Conditions in Elderly People: Machine Learning Approaches', *JMIR Medical Informatics*, 8(6), p. e16678. doi:

10.2196/16678.
Tax, N., Sidorova, N., Haakma, R. and van der Aalst, W.M., 2018. Mining Local Process Models with Constraints Efficiently: Applications to the Analysis of Smart Home Data. Medicine, 24, p.48.
Thirumalai, C., Chandhini, S. A. and Vaishnavi, M. (2017) 'Analysing the concrete compressive strength using Pearson and Spearman', in *2017 International conference of Electronics, Communication and Aerospace Technology (ICECA)*. IEEE, pp. 215–218. doi: 10.1109/ICECA.2017.8212799.
Tinetti, M. E. *et al.* (2014) 'Antihypertensive Medications and Serious Fall Injuries in a Nationally Representative Sample of Older Adults', *JAMA Internal Medicine*, 174(4), p. 588. doi: 10.1001/jamainternmed.2013.14764.
Tinker, A. (2002) 'The social implications of an ageing population', *Mechanisms of Ageing and Development*, 123(7), pp. 729–735. doi: 10.1016/S0047-6374(01)00418-3.
Travers, J. *et al.* (2019) 'Delaying and reversing frailty: a systematic review of primary care interventions', *British Journal of General Practice*, 69(678), pp. e61–e69. doi: 10.3399/bjgp18X700241.
Triki, S., Hanachi, C., Gleizes, M.P., Glize, P. and Rouyer, A., 2015. Modelling and Simulating Collaborative Scenarios for Designing an Assistant Ambient System that Supports Daily Activities. In Computational Collective Intelligence (pp. 191-202). Springer, Cham.
Tseng, T. and Flechner, L. (2011) 'Understanding results: P-values, confidence intervals, and number need to treat', *Indian Journal of Urology*, 27(4), p. 532. doi: 10.4103/0970-1591.91447.
Tukey, J. (1997) *Exploratory data analysis*.
U.S. Census Bureau (2019) *Quick Facts - Boston, Massachusetts*, *Online*. Available at: http://quickfacts.census.gov/qfd/states/25/2507000.html (Accessed: 20 December 2020).
U.S Department of Health & Human Services (2015) *International classification of diseases (ICD)*, *Online*. Available at: https://www.cdc.gov/nchs/icd/icd9.htm (Accessed: 21 December 2020).
United Nations (2017) *World population ageing 2017 - Highlights*, *Department of Economic and Social Affairs, Population Division*. Available at: https://www.un.org/en/development/desa/population/publications/pdf/ageing/WPA2017_Highlights.pdf.
Valero-Ramon, Z. *et al.* (2019) 'A Dynamic Behavioral Approach to Nutritional Assessment using Process Mining', in *2019 IEEE 32nd International Symposium on Computer-Based Medical Systems (CBMS)*. IEEE, pp. 398–404. doi: 10.1109/CBMS.2019.00085.
Vathy-Fogarassy, Á., Vassányi, I. and Kósa, I. (2022) 'Multi-level process mining methodology for exploring disease-specific care processes', *Journal of Biomedical Informatics*, 125. doi: 10.1016/j.jbi.2021.103979.
Verghese, J. *et al.* (2021) 'Trajectories of frailty in aging: Prospective cohort study', *PLOS ONE Journal*. Edited by A. Bayer, 16(7), p. e0253976. doi: 10.1371/journal.pone.0253976.
Virtanen, P. *et al.* (2020) 'SciPy 1.0: fundamental algorithms for scientific computing in Python', *Nature Methods*, 17(3), pp. 261–272. doi: 10.1038/s41592-019-0686-2.

Vitali, M. and Pernici, B., 2015. Pie-processes in Events: Interconnections in Ambient Assisted Living. In OTM Confederated International Conferences" On the Move to Meaningful Internet Systems" (pp. 157-166). Springer, Cham.

Walston, J. *et al.* (2006) 'Research agenda for frailty in older adults: Toward a better understanding of physiology and etiology: Summary from the American Geriatrics Society/National Institute on Aging research conference on frailty in older adults', *Journal of the American Geriatrics Society*, 54(6), pp. 991–1001. doi: 10.1111/j.1532-5415.2006.00745.x.

Weiskopf, N. G. and Weng, C. (2013) 'Methods and dimensions of electronic health record data quality assessment: Enabling reuse for clinical research', *Journal of the American Medical Informatics Association*, 20(1), pp. 144–151. doi: 10.1136/amiajnl-2011-000681.

Welstead, M. *et al.* (2020) 'A Systematic Review of Frailty Trajectories: Their Shape and Influencing Factors', *The Gerontologist*. Edited by P. C. Heyn. doi: 10.1093/geront/gnaa061.

Westergaard, D. *et al.* (2019) 'Population-wide analysis of differences in disease progression patterns in men and women', *Nature Communications*. Springer US, 10(1), pp. 1–14. doi: 10.1038/s41467-019-08475-9.

Wilson, D. *et al.* (2017) 'Frailty and sarcopenia: The potential role of an aged immune system', *Ageing Research Reviews*, 36, pp. 1–10. doi: 10.1016/j.arr.2017.01.006.

Wolff, J. L., Starfield, B. and Anderson, G. (2002) 'Prevalence, Expenditures, and Complications of Multiple Chronic Conditions in the Elderly', *Archives of Internal Medicine*, 162(20), p. 2269. doi: 10.1001/archinte.162.20.2269.

Wolf, H., Herrmann, K. and Rothermel, K., 2013. Dealing with uncertainty: Robust workflow navigation in the healthcare domain. ACM Transactions on Intelligent Systems and Technology (TIST), 4(4), p.65.

Wou, F. *et al.* (2013) 'The predictive properties of frailty-rating scales in the acute medical unit', *Age and Ageing*, 42(6), pp. 776–781. doi: 10.1093/ageing/aft055.

Wynn, M. T. and Sadiq, S. (2019) 'Responsible Process Mining - A Data Quality Perspective', *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 11675 LNCS, pp. 10–15. doi: 10.1007/978-3-030-26619-6_2.

Xu He, Fan Min and Zhu, W. (2014) 'Comparison of Discretization Approaches for Granular Association Rule Mining', *Canadian Journal of Electrical and Computer Engineering*, 37(3), pp. 157–167. doi: 10.1109/CJECE.2014.2343258.

Xue, Q.-L. (2011) 'The Frailty Syndrome: Definition and Natural History', *Clinics in Geriatric Medicine*, 27(1), pp. 1–14. doi: 10.1016/j.cger.2010.08.009.The.

Zaninotto, P. *et al.* (2020) 'Polypharmacy is a risk factor for hospital admission due to a fall: evidence from the English Longitudinal Study of Ageing', *BMC Public Health*, 20(1), p. 1804. doi: 10.1186/s12889-020-09920-x.

Zayas, C. E. *et al.* (2016) 'Examining Healthcare Utilization Patterns of Elderly Middle-Aged Adults in the United States.', *Proceedings of the ... International Florida AI Research Society Conference. Florida AI Research Symposium*, 2016, pp. 361–366. Available at: http://www.ncbi.nlm.nih.gov/pubmed/27430035.

Appendices

A.1 Certificate of CITI Program Completion for MIMIC-III Dataset

8.6.1 Part A

COLLABORATIVE INSTITUTIONAL TRAINING INITIATIVE (CITI PROGRAM)

COMPLETION REPORT - PART 1 OF 2
COURSEWORK REQUIREMENTS*

* NOTE: Scores on this Requirements Report reflect quiz completions at the time all requirements for the course were met. See list below for details. See separate Transcript Report for more recent quiz scores, including those on optional (supplemental) course elements.

- **Name:** Nik Fatinah N Mohd Farid (ID: 5816966)
- **Email:** scnfnm@leeds.ac.uk
- **Institution Affiliation:** Massachusetts Institute of Technology Affiliates (ID: 1912)
- **Institution Unit:** Computing

- **Curriculum Group:** Human Research
- **Course Learner Group:** Data or Specimens Only Research
- **Stage:** Stage 1 - Basic Course

- **Report ID:** 20852704
- **Completion Date:** 19-Sep-2016
- **Expiration Date:** 19-Sep-2019
- **Minimum Passing:** 90
- **Reported Score*:** 92

REQUIRED AND ELECTIVE MODULES ONLY	DATE COMPLETED	SCORE
Belmont Report and CITI Course Introduction (ID: 1127)	17-Sep-2016	3/3 (100%)
History and Ethics of Human Subjects Research (ID: 498)	14-Sep-2016	5/7 (71%)
Basic Institutional Review Board (IRB) Regulations and Review Process (ID: 2)	16-Sep-2016	5/5 (100%)
Records-Based Research (ID: 5)	17-Sep-2016	3/3 (100%)
Genetic Research in Human Populations (ID: 6)	17-Sep-2016	5/5 (100%)
Populations in Research Requiring Additional Considerations and/or Protections (ID: 16680)	18-Sep-2016	5/5 (100%)
Research and HIPAA Privacy Protections (ID: 14)	18-Sep-2016	4/5 (80%)
Conflicts of Interest in Research Involving Human Subjects (ID: 488)	19-Sep-2016	5/5 (100%)
Massachusetts Institute of Technology (ID: 1290)	19-Sep-2016	No Quiz

For this Report to be valid, the learner identified above must have had a valid affiliation with the CITI Program subscribing institution identified above or have been a paid Independent Learner.

Verify at: https://www.citiprogram.org/verify/?68580891-d191-4ec7-8f61-6e3487b7b972

CITI Program
Email: support@citiprogram.org
Phone: 888-529-5929
Web: https://www.citiprogram.org

8.6.2 Part B

COLLABORATIVE INSTITUTIONAL TRAINING INITIATIVE (CITI PROGRAM)

COMPLETION REPORT - PART 2 OF 2
COURSEWORK TRANSCRIPT**

** NOTE: Scores on this Transcript Report reflect the most current quiz completions, including quizzes on optional (supplemental) elements of the course. See list below for details. See separate Requirements Report for the reported scores at the time all requirements for the course were met.

- **Name:** Nik Fatinah N Mohd Farid (ID: 5816966)
- **Email:** scnfnm@leeds.ac.uk
- **Institution Affiliation:** Massachusetts Institute of Technology Affiliates (ID: 1912)
- **Institution Unit:** Computing

- **Curriculum Group:** Human Research
- **Course Learner Group:** Data or Specimens Only Research
- **Stage:** Stage 1 - Basic Course

- **Report ID:** 20852704
- **Report Date:** 19-Sep-2016
- **Current Score**:** 92

REQUIRED, ELECTIVE, AND SUPPLEMENTAL MODULES	MOST RECENT	SCORE
History and Ethics of Human Subjects Research (ID: 498)	14-Sep-2016	5/7 (71%)
Belmont Report and CITI Course Introduction (ID: 1127)	17-Sep-2016	3/3 (100%)
Records-Based Research (ID: 5)	17-Sep-2016	3/3 (100%)
Genetic Research in Human Populations (ID: 6)	17-Sep-2016	5/5 (100%)
Research and HIPAA Privacy Protections (ID: 14)	18-Sep-2016	4/5 (80%)
Conflicts of Interest in Research Involving Human Subjects (ID: 488)	19-Sep-2016	5/5 (100%)
Basic Institutional Review Board (IRB) Regulations and Review Process (ID: 2)	16-Sep-2016	5/5 (100%)
Populations in Research Requiring Additional Considerations and/or Protections (ID: 16680)	18-Sep-2016	5/5 (100%)
Massachusetts Institute of Technology (ID: 1290)	19-Sep-2016	No Quiz

For this Report to be valid, the learner identified above must have had a valid affiliation with the CITI Program subscribing institution identified above or have been a paid Independent Learner.

Verify at: https://www.citiprogram.org/verify/?68580891-d191-4ec7-8f61-6e3487b7b972

Collaborative Institutional Training Initiative (CITI Program)
Email: support@citiprogram.org
Phone: 888-529-5929
Web: https://www.citiprogram.org

A.2 Letter of Access for Bradford Dataset

Bradford Teaching Hospitals NHS
NHS Foundation Trust

Bradford Institute for Health Research
Temple Bank House
Bradford Royal Infirmary
Duckworth Lane
Bradford
West Yorkshire
BD9 6RJ
Tel: 01274 383418

18th May 2017

Nik Farid

BY EMAIL

Dear Nik,

LETTER OF ACCESS FOR RESEARCH

Study: Born in Bradford
BTHFT Local Reference No: 885
PI: Prof John Wright

If you have not already provided your ID documentation to the R&D Office 1 please do so on the day you start work at Bradford – this is a term of your Letter of Access

This letter should be presented to each participating organisation before you commence your research at that site.

In accepting this letter, each participating organisation confirms your right of access to conduct research through their organisation for the purpose and on the terms and conditions set out below. This right of access commences on 18/05/2017 and ends on 31/08/2020 unless terminated earlier in accordance with the clauses below.

You have a right of access to conduct such research as confirmed in writing in the letter of permission for research from Bradford Teaching Hospitals NHS Foundation Trust. Please note that you cannot start the research until the Principal Investigator for the research project has received a letter from us giving confirmation from the individual organisation of their agreement to conduct the research.

The information supplied about your role in research at the organisation has been reviewed and you do not require an honorary research contract with the organisation. We are satisfied that such pre-engagement checks as we consider necessary have been carried out. Evidence of checks should be available on request to the organisation.

A.3 List of frailty deficits based on Pareto Principle selection within different frailty stages

Not frail > Mild		Mild > Moderate		Moderate > Severe		Severe	
Deficit (Total count: 5,5,35)	**# Occ. (%)**	**Deficit (Total count: 6,374)**	**# Occ. (%)**	**Deficit (Total count: 2,978)**	**# Occ. (%)**	**Deficit (Total count: 881)**	**# Occ. (%)**
Polypharmacy	996 (18.0)	Polypharmacy	834 (13.1)	Housebound	198 (6.6)	Fall	57 (6.5)
Hypertension	434 (7.8)	Housebound	388 (6.1)	Anaemia haematinic deficiency	184 (6.2)	Atrial fibrillation	53 (6.0)
Diabetes	305 (5.5)	Urinary system disease	329 (5.2)	Urinary system disease	183 (6.1)	Housebound	52 (5.9)
Respiratory disease	299 (5.4)	Atrial fibrillation	310 (4.9)	Atrial fibrillation	171 (5.7)	Heart failure	49 (5.6)
Urinary system disease	287 (5.2)	Fall	293 (4.6)	Polypharmacy	155 (5.2)	Memory cognitive problem	49 (5.6)
Ischaemic heart disease	274 (5.0)	Chronic kidney disease	291 (4.6)	Fall	152 (5.1)	Anaemia haematinic deficiency	45 (5.1)
Visual impairment	248 (4.5)	Anaemia haematinic deficiency	282 (4.4)	Memory cognitive problem	150 (5.0)	Requirement for care	43 (4.9)
Housebound	236 (4.3)	Heart failure	241 (3.8)	Heart failure	140 (4.7)	Hypotension	42 (4.8)
Atrial fibrillation	210 (3.8)	Visual impairment	219 (3.4)	Chronic kidney disease	132 (4.4)	Mobility transport problem	40 (4.5)
Arthritis	207 (3.7)	Dyspnoea	213 (3.3)	Hypotension	130 (4.4)	Social vulnerability	38 (4.3)
Anaemia haematinic deficiency	190 (3.4)	Respiratory disease	213 (3.3)	Requirement for care	109 (3.7)	Urinary system disease	37 (4.2)
Dyspnoea	187 (3.4)	Memory cognitive problem	208 (3.3)	Mobility transport problem	100 (3.4)	Osteoporosis	34 (3.9)
Fall	165 (3.0)	Hypertension	203 (3.2)	Dyspnoea	95 (3.2)	Cerebrovascular disease	31 (3.5)
Chronic kidney disease	140 (2.5)	Cerebrovascular disease	196 (3.1)	Fragility fracture	93 (3.1)	Urinary incontinence	31 (3.5)
Cerebrovascular disease	132 (2.4)	Hypotension	179 (2.8)	Social vulnerability	90 (3.0)	Dyspnoea	25 (2.8)
Memory cognitive problem	109 (2.0)	Ischaemic heart disease	178 (2.8)	Cerebrovascular disease	87 (2.9)	Respiratory disease	25 (2.8)
Heart failure	108 (2.0)	Mobility transport problem	174 (2.7)	Osteoporosis	82 (2.8)	Fragility fracture	24 (2.7)
Hypotension	101 (1.8)	Arthritis	159 (2.5)	Visual impairment	68 (2.3)	Ischaemic heart disease	23 (2.6)

A.4 Experiment #1: MIMIC-III Dataset

UNIVERSITY OF LEEDS School of Computing	**EXPERIMENT DOCUMENTATION** **Experiment title:** Approachability of process mining into MIMIC-III dataset **Researcher's name:** Nik Farid	**Date of experiment** 26 March 2018 **Experiment code** M3_APM_C4E1

Area of investigation

This experiment is data profiling of Bradford Dataset as an initial step for determining features for confirmatory analysis

Data source

The MIMIC-III dataset

Research question

1. What is the most followed path and the exceptional path of the frail elderly patient?
2. Is there any difference in path followed by different frailty categories?
3. Is there any bottleneck in the path of frail elderly patient?

Hypothesis

Process mining can be applied into frail elderly MIMIC-III dataset to analyse the hospital workflow

Method:

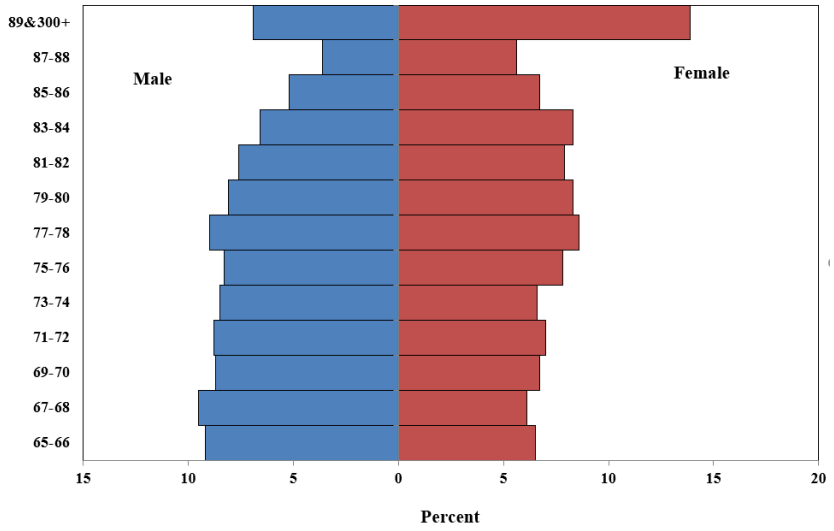
The general method used is the Process Mining Project Methodology (PM2) which includes Stage I: Planning, Stage II: Extraction, Stage III: Data Transformation and Loading and Stage IV: Mining and Analysis and Stage V: Evaluation.

Stage I: Planning
In this stage three research questions were addressed to investigate the hospital workflow within the frail elderly patients. The questions were based on the frequently posed by the healthcare professionals following (Mans *et al.*, 2013; Rojas, Arias and Sepúlveda, 2015).

Stage II: Extraction
The extraction was done from the raw database to only include elderly patient. The elderly patient is selected based on the age of 65 years from the first admission into hospital. The query to extract elderly patient with minimum aged of 89 years old is as followed:

```
WITH first_admission_time AS
(
  SELECT
      p.subject_id, p.dob, p.gender
      , MIN (a.admittime) AS first_admittime
      , MIN( ROUND( (cast(admittime as date) - cast(dob as date)) / 365.242,2) )
          AS first_admit_age
  FROM mimiciii.patients p
  INNER JOIN mimiciii.admissions a
  ON p.subject_id = a.subject_id
  GROUP BY p.subject_id, p.dob, p.gender
  ORDER BY p.subject_id
)
SELECT subject_id, dob, gender, first_admittime, first_admit_age
FROM first_admission_time
WHERE first_admit_age >87
ORDER BY subject_id
```

The distribution of selected elderly patient is illustrated using the pyramid population as follows:



Stage III: Data Transformation and Loading

The transformation is done in several steps on the extracted dataset. It aims in creating an event log which incorporates the transactional information. It follows the PM2 processing steps starts with (i) filtering, (ii) creating views on event log and (iii) enriching logs. The details of data transformation steps are presented in table with percentage calculated from the total row of extracted data.

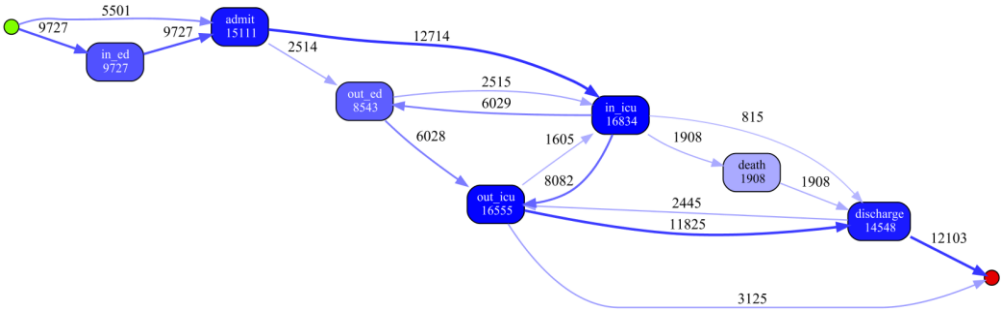
Step	# rows	Percentage (%)
(0) Extracted Log	71,511,683	100.00
(1) Creating Views	2,483,439	3.47
(2) Filtering Log	188,804	0.26
(3) Enriching Log:	Frailty Category	
Mild	63,049	< 0.0001
Moderate	34,433	< 0.0001
Severe	7,077	< 0.0000001

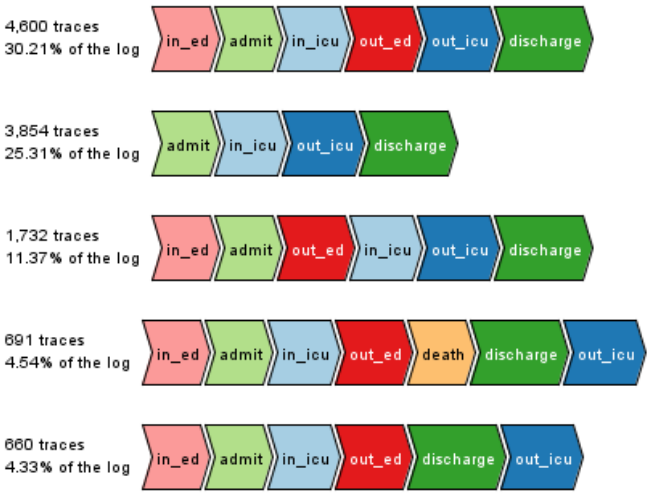
Results

Stage IV: Mining and Analysis

Result and discussion
The main approach applied to answer questions in the experiment is using the Directly Follows visual Miner (DFvM) plugin in ProM. It is an extension of the previous plug-in in the ProM, Inductive Visual Miner (IvM). The plugin produces a directly follows model. Activities are represented by blue rectangles, and edge represents from an activity a to activity b if a is followed directly. The darker colour of activity box and lines is proportional to the frequency observed in the log. The experiment used the default setting of the plugin.

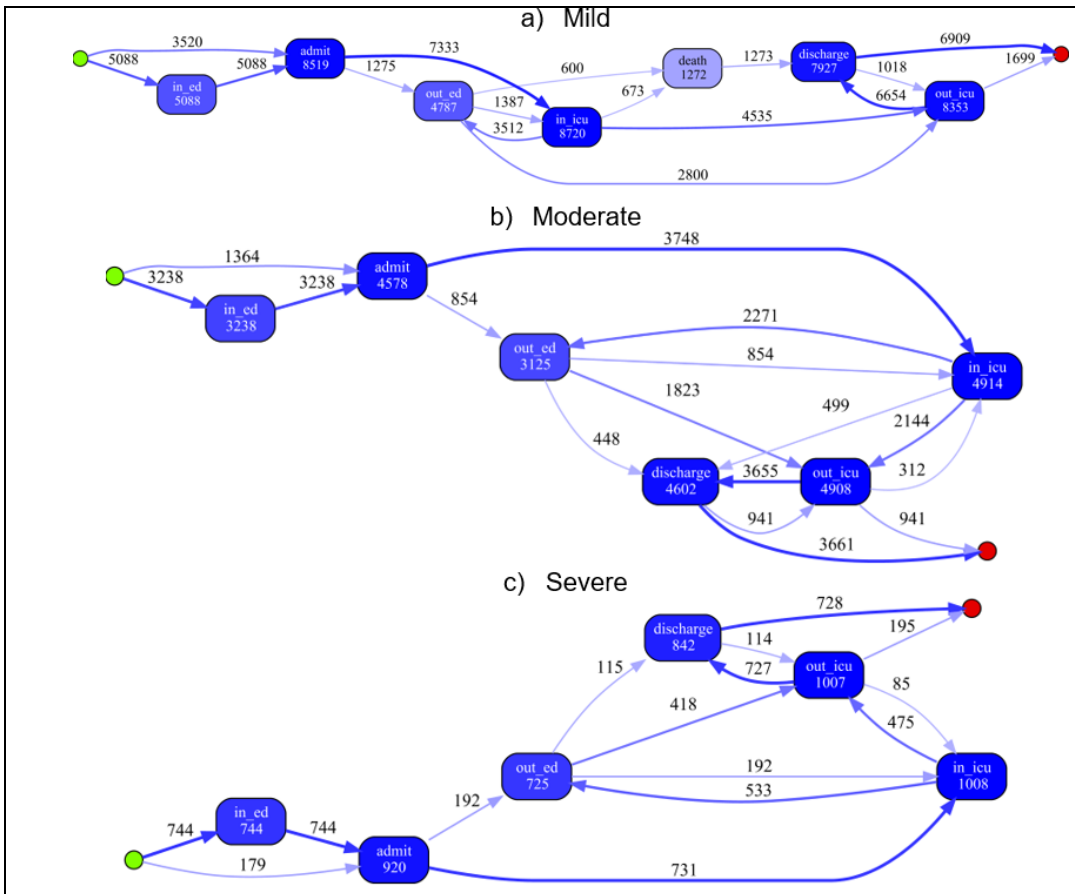
RQ1: The model produced using the DFvM plugin in ProM as an approach to address RQ1. Other than that, top 5 trace variants are presented in the following Figure using ProM to identify the most followed path by the frail elderly patient. The model illustrates the common workflow within the hospital for frail elderly patient.





Finding RQ1: The pathway of hospital admission could either end with discharge from hospital or out of the ICU. The top 5 trace variants out of 100 total variants covered 11,538 (75.8%) of all traces. The flow of admission could start from registering to the ED or direct to hospital admission. A compelling pattern was observed that some patients could be out of the ICU after been discharged from the hospital (from fourth and fifth trace variants).

RQ2: The second question of the experiment is looking into the pathway of each frailty categories and determine any similarities or dissimilarities between process models. The same approach following the previous question with process tree as the visualisation using the IvM plugin.



Finding RQ2: Process models from above figure are generated using the previous approach. They undergo visual evaluation to identify differences and similarities between them for comparison purpose.

Similarity:

i) The activities admit -> discharge and in_icu -> out_icu happened in sequential order for all categories

ii) The processes could end by going through the discharge activity or out of the ICU activity

Differences:

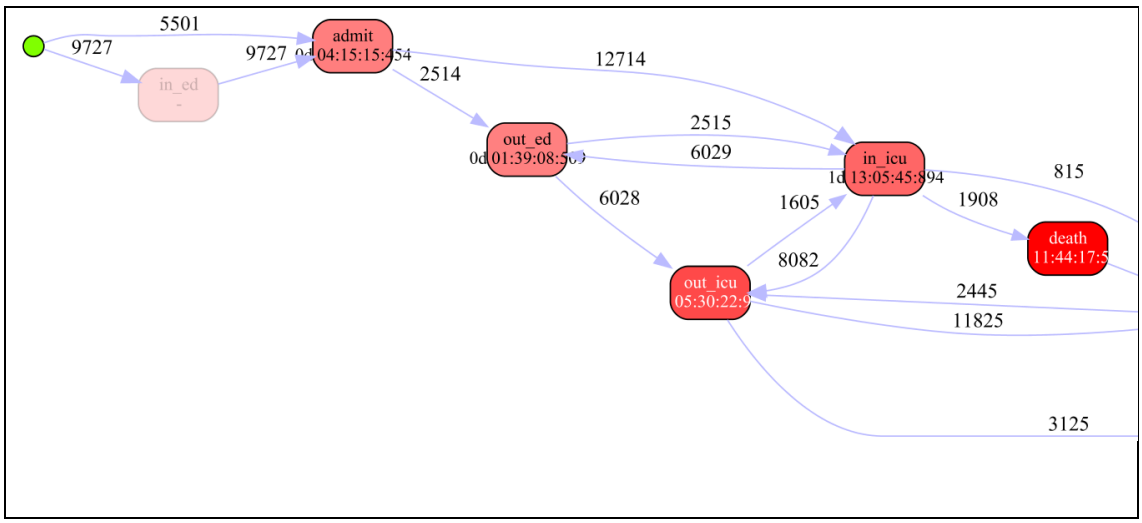
i) In moderate and severe the death activity can happen before the activities out of the ICU or discharge

ii) In mild no death activity present unlike in moderate and severe. It makes the activity death in mild not so significant as in the other models

iii) In severe most cases 744 (81%) start the process by going into the emergency department (ED) before admitting to hospital or ICU. Unlike mild and moderate, the process can either start with admission into hospital or ICU

RQ3: The last question of the experiment is to identify the bottleneck of the hospital admission process of frail patient. The directly-follows model from the RQ1

Performance analysis based on the similarity comparison made from previous RQ.


5501
9727
admit
9727
12714
2514
in_ed
out_ed
2515
6029
in_icu
815
1908
death
6028
1605
8082
out_icu
2445
11825
3125

A.5 Experiment #2: MIMIC-III Dataset (Trajectories)

UNIVERSITY OF LEEDS School of Computing	**EXPERIMENT DOCUMENTATION** **Experiment title:** Frailty Progression within Different Categories **Researcher's name:** Nik Farid	**Date of experiment** 16 June 2018 **Experiment code** M3_FPDF_C4E2

Area of investigation

This experiment is analysing the variability of frailty progression within frailty categories

Data source

The dataset is a subset of MIMIC-III database of frail elderly patient

Research question

Can process mining be used to explore the variation in frailty deficit pathway within frailty categories?

Hypothesis

Process mining can be used to explore the variation in frailty deficit pathways within frailty categories

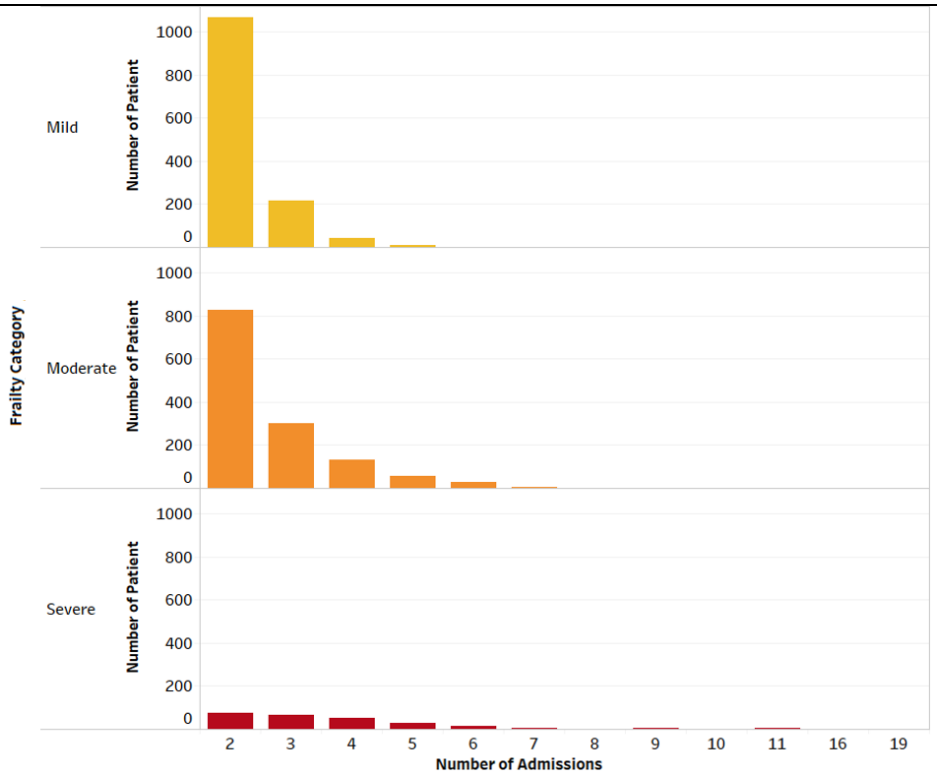
Method

1) **Extract and transform** steps are done in M3_APM_C4E1. A selection criterion is added in this experiment where patient with minimum admission of two is only included.
2) **Load** of the extracted data into Python Jupyter, Disco and ProM for analysis

Results and Discussion of Mining and Analysis

1) The distribution of patient with the selection criterion is illustrated with bar chart as follows:

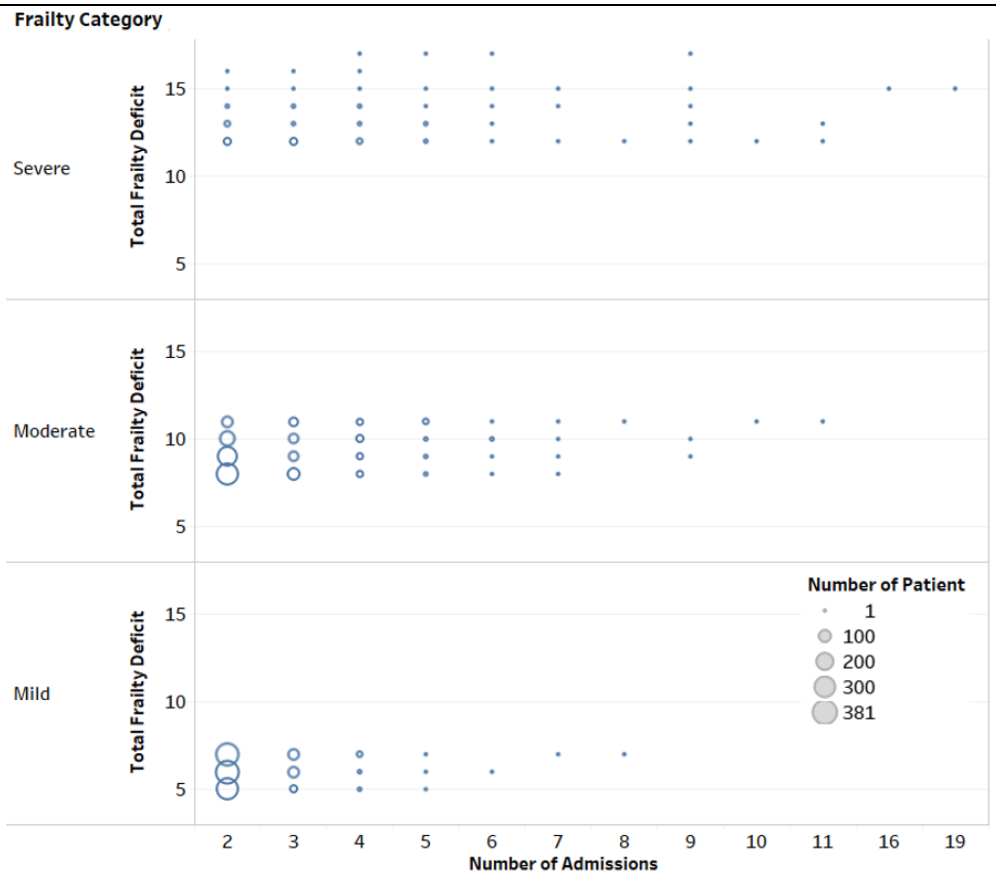
   About 3,446 (17.6%) patient matched with the selection criterion from the initial dataset. A total of 2,950 (85.6%) are frail elderly patient where most of them 2,934 (85.1%) had number of admissions between two to eight and only 16 (0.5%) had number of admissions more than eight.



Number of Admissions

Frailty Category	2	3	4	5	6	7	8	9	10	11	16	19	Grand Total
Mild	1,064	218	43	11	1	2	1						**1,340**
Moderate	828	301	131	56	26	6	1	2	2	1			**1,354**
Severe	75	68	50	27	16	7	2	5	1	3	1	1	**256**
Grand Total	**1,967**	**587**	**224**	**94**	**43**	**15**	**4**	**7**	**3**	**4**	**1**	**1**	**2,950**

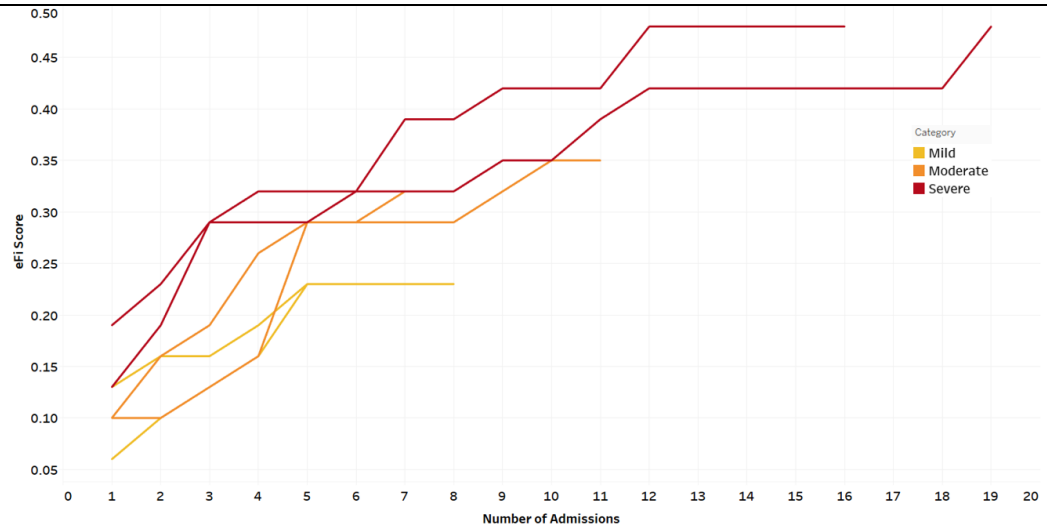
2) Number of admission VS number of cumulative frailty deficit within different categories



The figure shows the distribution of patient admission and number of cumulative deficits of latest admission. The y-axis represents the total number of cumulative deficits where x-axis represents the total number of admissions. The label of vertical rows shows the frailty categories based on patients' cumulative deficits. The size of the circle in the figure indicates the number of patients with their admission and cumulative deficit respectively.

Observation: Most of the patient 2,554 (86.5%) had number of admissions of two and three. The total cumulative frailty deficits cut off point is following the domain expert which creates a significant line between frailty categories. The progressed frailty categories are highly likely to have higher number of admissions.

3) Randomly select two (2) patient from each category is done to illustrates that variation of frailty progression is present

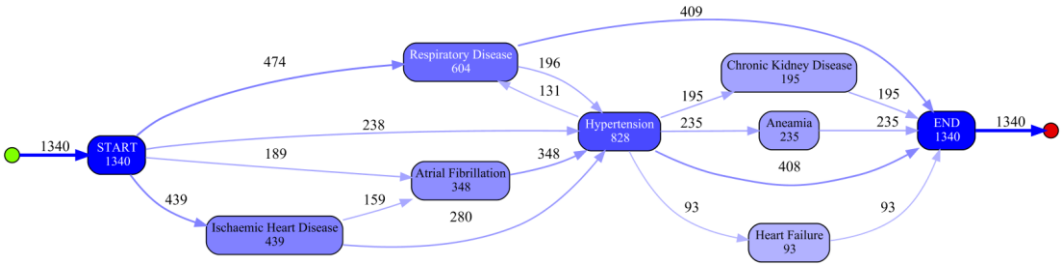


The step graph illustrates the progression of frailty varies between categories of randomly selected six patients. The x-axis represents the eFi score calculated at each admissions and y-axis represents the number of admissions.

4) Frailty trajectories or graph network illustrating the flow of frailty deficit starting from first until final deficits created using process mining approach. Only the first occurrence of the deficit event is considered for the trajectories

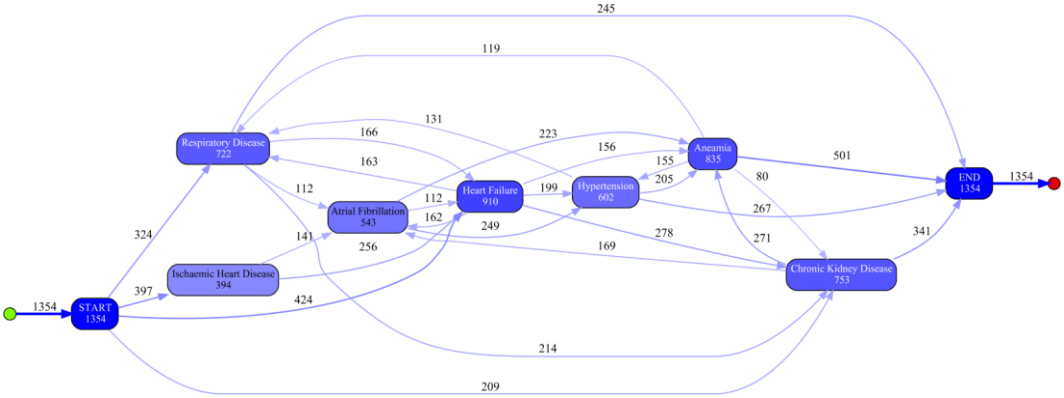
Frailty flow/trajectories:

i. Mild



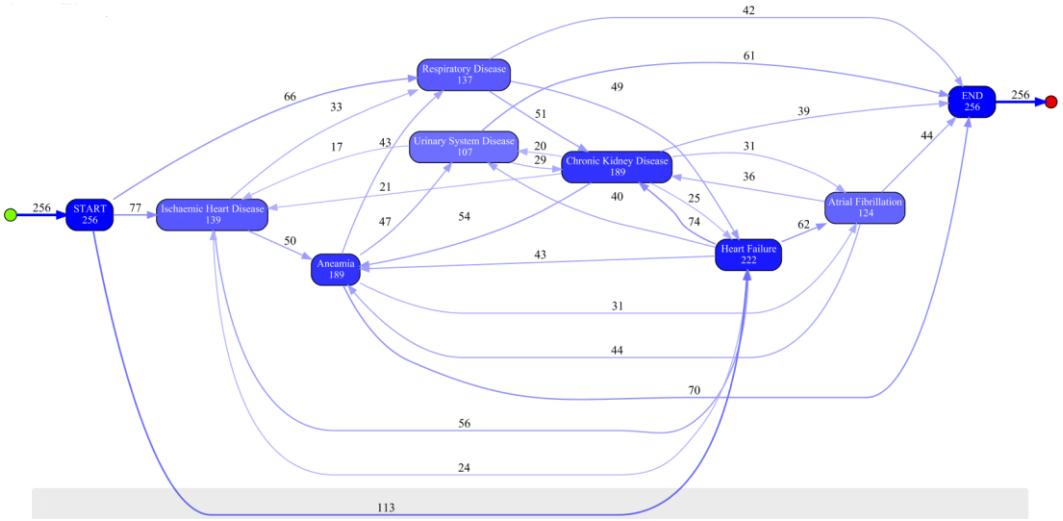
(25% of most frequent activities and 1% of the most occurring paths) (Fitness:0.9678)

ii. Moderate



(25% of most frequent activities and 1% of the most occurring paths) (Fitness: 0.97735)

iii. Severe



(25% of most frequent activities and 1% of the most occurring paths) (Fitness: 0.97268)

Finding: The process models generated using the Directly Followed visual Miner (DFvM) plugin in ProM. All models are set to show 25% of most frequent activities and 1% of most occurring path.

Overview of the event logs of each category:

Category	# Case	# events	Events per cases	Case duration	Total variants
Mild	1,340	8,137	Min 5, mean 6 and max 7	Median: 4 months and 2 weeks Mean: 1 year 4 months 2 weeks 4 days	1,337 (99.78%)*
Moderate	1,354	12,403	Min 8, mean 9, max 11	Median: 10 months and 3 weeks Mean: 1 year 10 months 1 week and 5 days	1,354 (100.0%)*
Severe	256	3,306	Min 12, mean 13, max 17	Median: 1 year and 10 months Mean: 2 years 8 months 1 week and 5 days	256 (100.0%)*

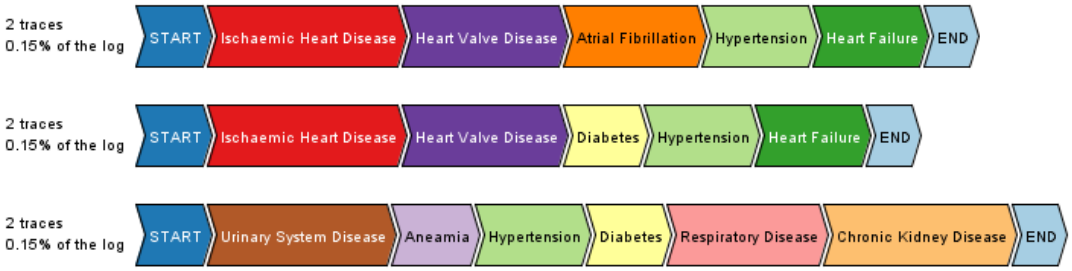
*Variation percentage (Number of trace variant / total case) * 100

Top five activities (deficit associated with frailty) most occur within each category:

Category	1st activity	2nd activity	3rd activity	4th activity	5th activity
Mild	Hypertension, 828 (10.2%)	Heart Failure, 815 (10.0%)	Respiratory disease, 734 (9.0%)	Ischaemic heart disease, 711 (8.7%)	Chronic kidney disease, 692 (8.5%)
Moderate	Heart failure, 1,116 (9.0%)	Chronic kidney disease, 1,045 (8.4%)	Anaemia, 987 (8.0%)	Respiratory disease, 986 (8.0%)	Ischaemic heart disease, 926 (7.5%)
Severe	Chronic kidney disease, 244 (7.4%)	Heart failure, 243 (7.3%)	Anaemia, 236 (7.1%)	Respiratory disease, 216 (6.5%)	Ischaemic heart disease, 216 (6.5%)

5) Most common variants within frailty categories

Mild:



Finding: The total number of variants are illustrated in Table overview of the event logs for each category. Only mild category has common variant between traces, while the other two categories all has its own unique traces.

Conclusion

The experiment shows that routinely collected data from the patient record can be used for process mining analysis in identifying the variation in frailty trajectories. It presents the variations hidden between frailty categories that is highly linked with frailty deficits. However, future work is highly recommended in exploring the potential causal link between deficits to understand the frailty progression.

A.6 Experiment #6: Bradford Dataset (Frailty Progression)

UNIVERSITY OF LEEDS School of Computing	**EXPERIMENT DOCUMENTATION** **Experiment title:** Frailty Progression and Its Association with Focused Frailty Deficits **Researcher's name:** Nik Farid	**Date of experiment** 17 November 2020 **Experiment code** BD_FPAD_C6E5

Area of investigation

This experiment is determining the association of focused frailty deficits within frailty stages

Data source

The Bradford dataset

Research question

RQ #1: Can process mining discover and quantify the variability in frailty progression based on the status of deficits of concern?
RQ #2: Is it possible to uncover the differences in sequence of deficits of concern using process mining?
RQ #3: Can process mining determine and evaluate the differences between patterns of concerns?

Hypothesis

The discovered frailty progression model and its association with fall, hypertension and polypharmacy can be used as initial step in analysing the relatedness with frailty severity utilising process mining approach.

Method:

The general method used is the Process Mining Project Methodology (PM2) which includes Stage I: Planning, Stage II: Extraction, Stage III: Data Transformation and Loading and Stage IV: Mining and Analysis and Stage V: Evaluation.

I: Planning - Three research questions are addressed (as above) to investigate the frailty progression within deficits of concern. The questions are based on the aim to determine the association of deficits of concern within frailty stages.

II: Extraction – Minimum aged of 65 years elderly patient at the start of study duration within year 2003 – 2018. Three additional exclusion are:

i. Patient has at least one year record within study duration
ii. Patient aged over 85 years old who their final records in either fit or mild category
iii. Maximum average accumulation of frailty deficits is only three

A total of 12,395 patients with three deficits of concern (fall, hypertension, and polypharmacy) is 8,547 and without is 3,848.

III: Data Transformation and Loading – Creating view based on the structure of the data and research aim. Additional transformation steps are as follows:

i. Identification of frailty score – the first occurrence of each deficits associated with frailty was identified at each visit to GP practice
ii. Log enriching – additional events are extended into the initial event log which are frailty stages; fit, mild, moderate, and severe.

iii. Securing the events sequence – the order of the events must always start with frailty stage event, if frailty stage event and frailty deficits events shared similar timestamp.

Once events logs are created for sub-cohort with deficits of concern and without, save them in .csv file. Then load it into Disco and extract .XES files to load into ProM tool.

Results

Stage IV: Mining and Analysis

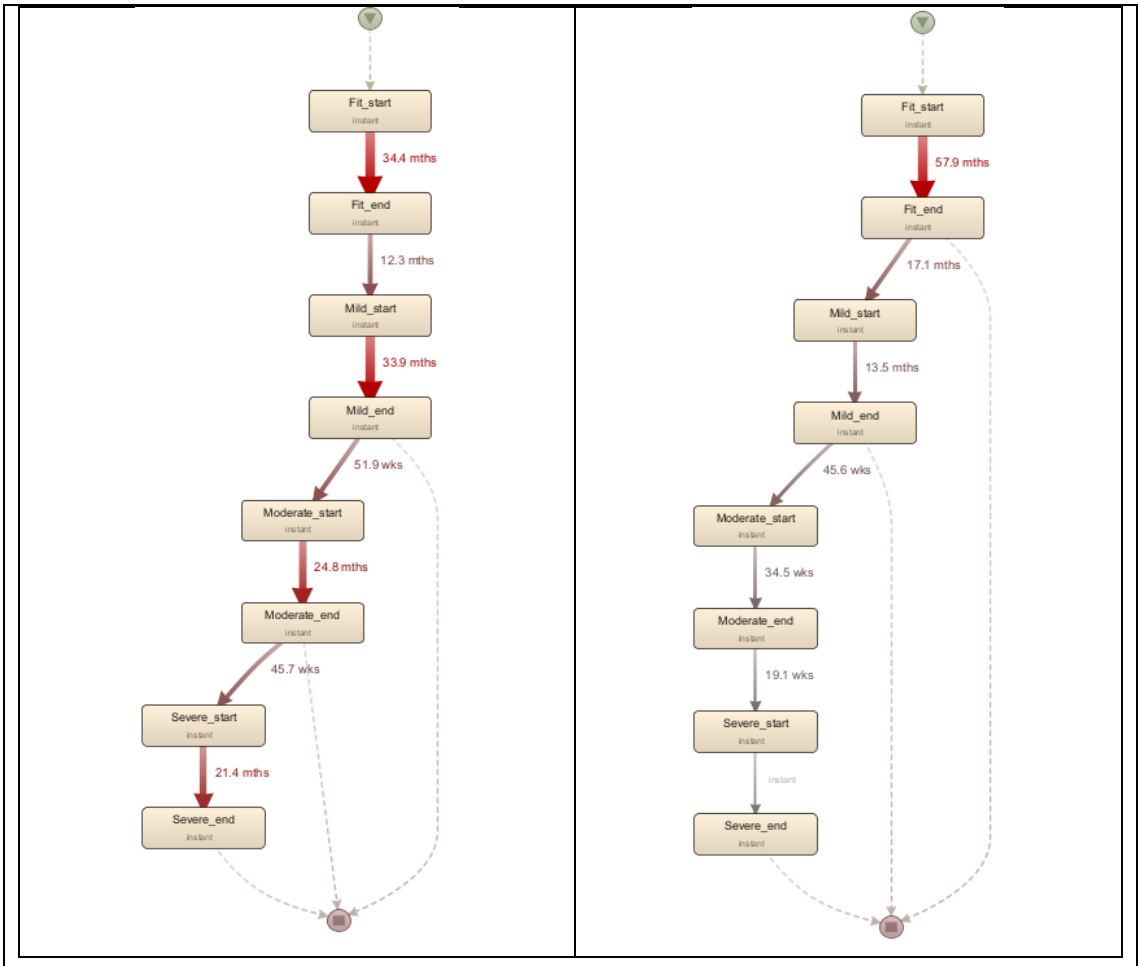
Result and discussion

1. The two sub-cohorts of with and without deficits of concerns are being analysed.

Descriptive statistics of two sub-cohorts:

	With Focused Deficits	**Without Focused Deficits**	**P-Value**
# patients	8,547	3,848	
# events	30,754	5,385	
Events per patient	3.6 (~4)	1.4 (~1)	
Trace Fitness	0.96	1.00	
Precision	1.00	1.00	
Generalisation	0.77	0.53	
Frailty Stages	Median Duration (IQR) in months		
Fit	34.4 (13.3 - 47.1) N = 8,547	**57.9 (25.0 - 81.9)** N = 3,848	**0.00**
Mild	**33.9 (14.3 - 47.8)** N = 8,514	13.5 (0 - 18.6) N = 1,432	**0.00**
Moderate	**24.8 (5.8 - 36.3)** N = 7,023	8.0 (0.0 – 9.9) N = 101	**0.00**
Severe	21.4 (0.0 – 33.7) N = 3,335	0.0 (0.0 – 0.0) N = 2	0.25
Transition Point	Median Duration (IQR) in months		
1	7.0 (2.1 – 16.6) N = 8,514	**9.0 (2.6 – 23.6)** **N = 1,432**	**0.00**
2	6.7 (2.0 – 16.5) N = 7,023	5.5 (1.7 – 14.6) N = 101	0.32
3	6.0 (1.9 – 14.6) N = 3,335	4.4 (3.6 – 5.2) N = 2	0.49

Activities 100% Path 100%	
With Focused Deficits	**Without Focused Deficits**
Number of patients at each frailty stages	
# Stage I: 8,547 # Stage II: 8,514 # Stage III: 7,023 # Stage IV: 3,335	# Stage I: 3,848 # Stage II: 1,432 # Stage III: 101 # Stage IV: 2

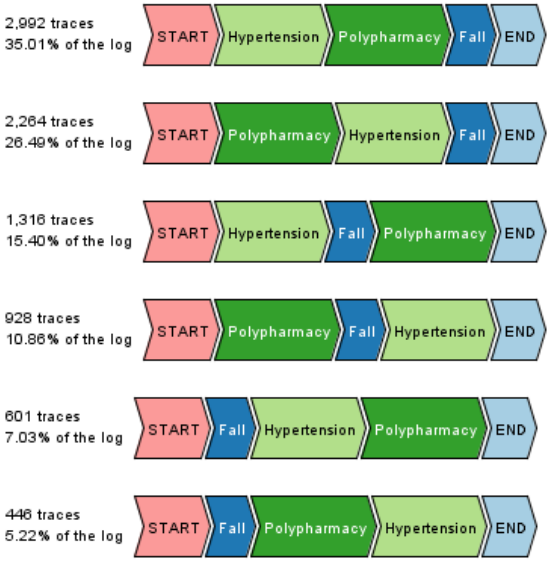


Notes on important points:

i. There is difference exists between the two sub-cohorts.

ii. The sub-cohort with deficits of concern are experiencing longer case duration compared to another sub-cohort

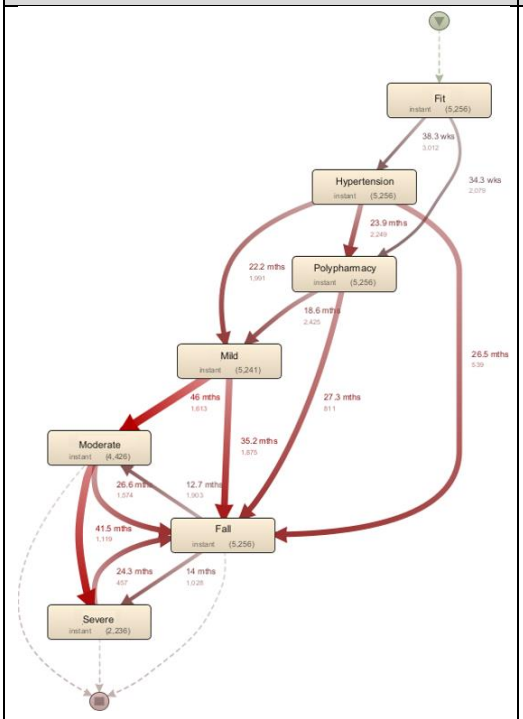
2. The association between deficits of concern were examined by implementing trace variant analysis of the log



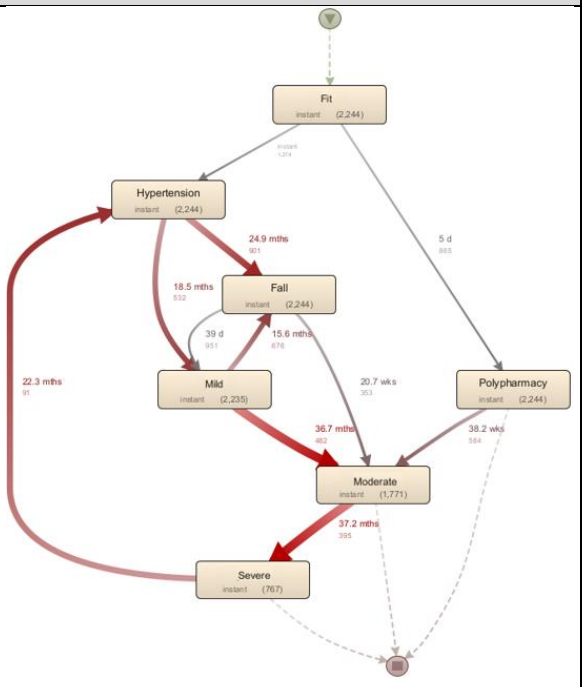
The descriptive statistics based on the Pattern of Sequence observed:

Pattern of Sequence	# Cases	Case Portion	Mean Case duration	Median Case Duration [IQR]
I	5,256	F: 60 Mi: 815 Mo: 2,190 Se: 2,236	10y, 6m	11y,1m [8y,1m – 13y,6m]
II	2,244	F: 9 Mi: 464 Mo: 1,004 Se: 767	9y, 10m	10y,2m [7y,2m – 12y,9m]
III	1,047	F: 9 Mi: 212 Mo: 494 Se: 332	9y, 6m	10y,0m [6y,6m – 12y,5m]

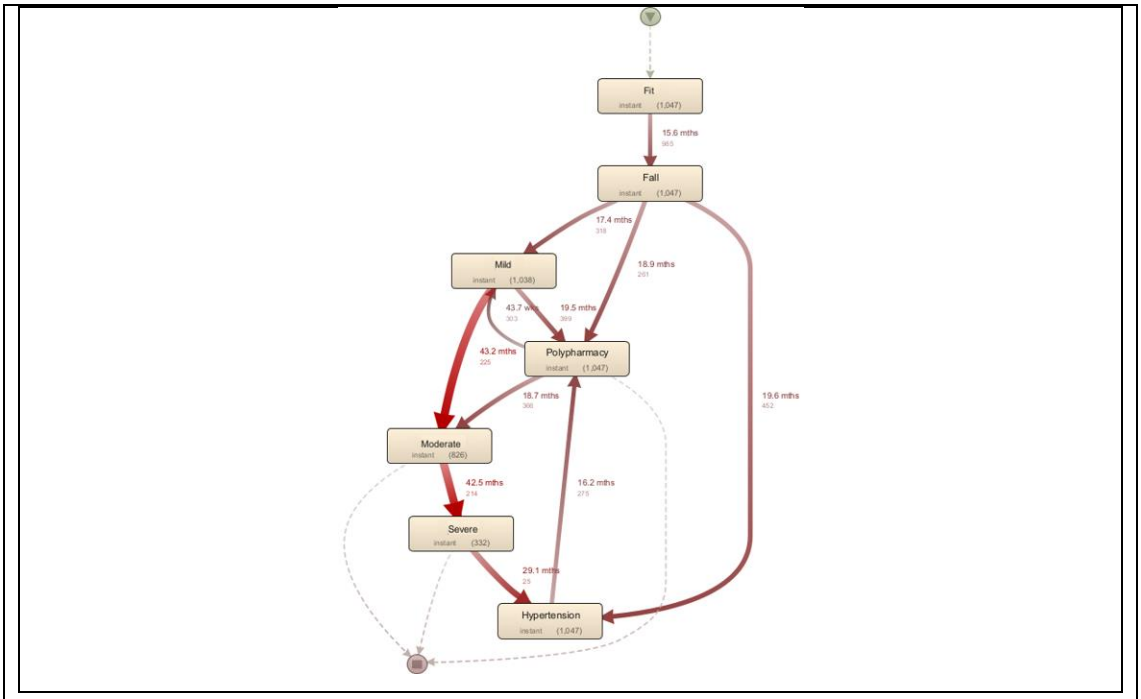
Pattern of Sequence I



Pattern of Sequence II



Pattern of Sequence III

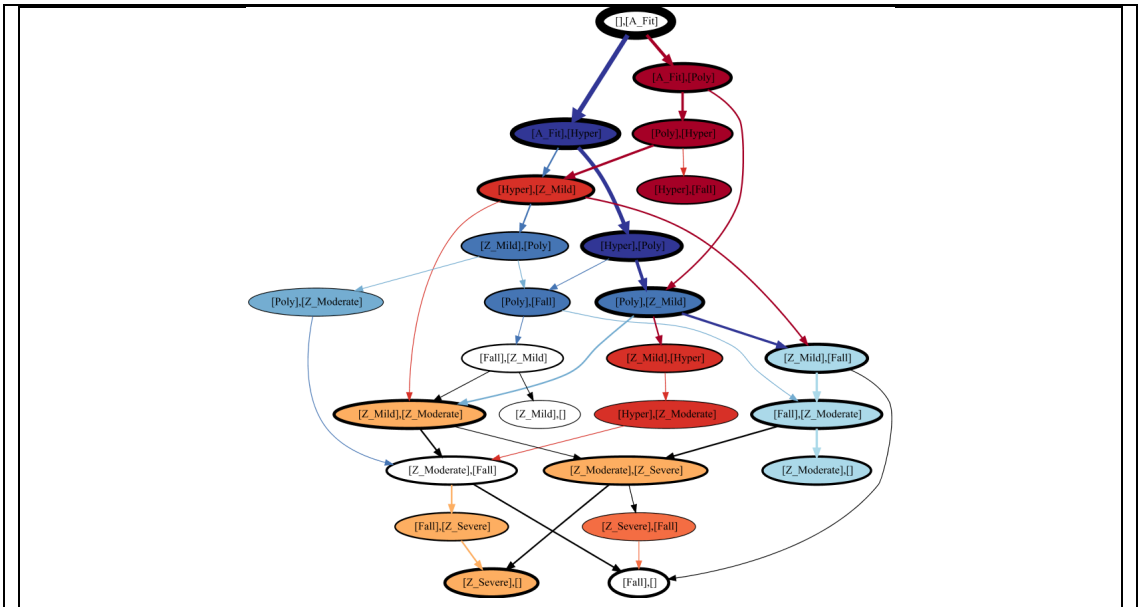


The dominant pattern of sequence is being analysed further.
Descriptive statistics of pattern of sequence I formed from Trace Variant I and II

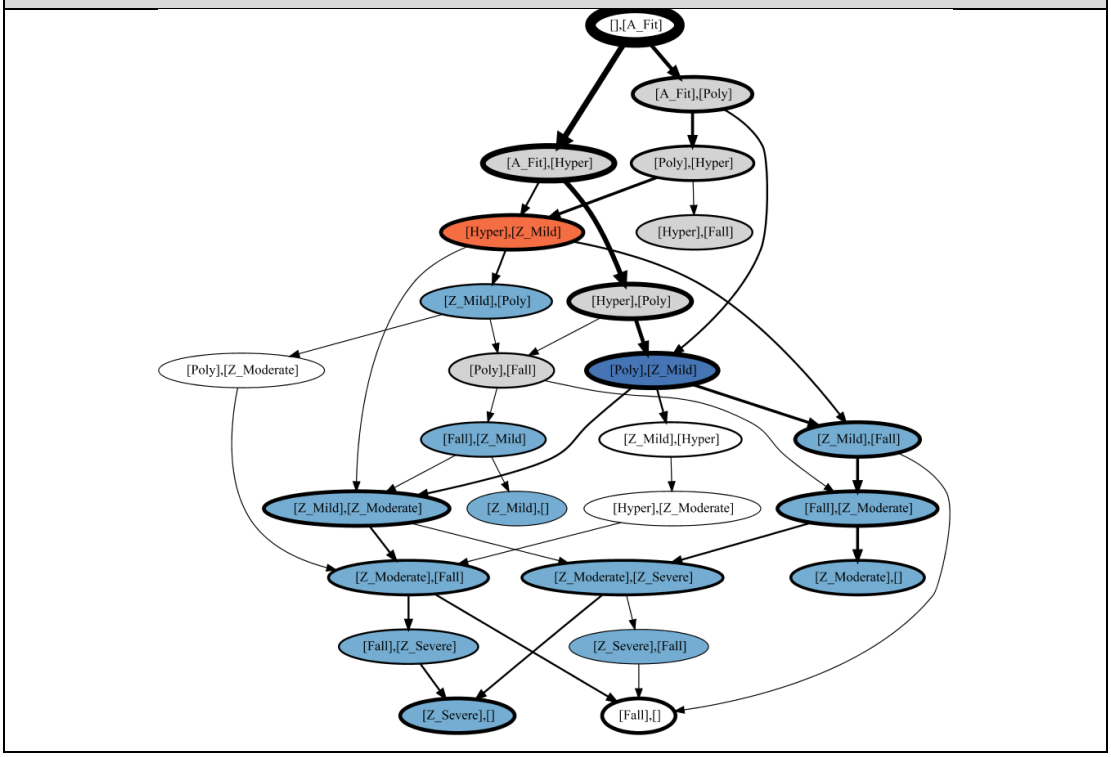
Characteristics	Variant I	Variant II
# Patient (%)	2,564 (58.5)	1,821 (41.5)
# Events (%)	15,817 (57.7)	11,581 (42.3)
# Trace Variant	29	33
# Activities	7	7
Activities Per Patient, Mean [Min -Max]	6 [4 – 7]	6 [4 – 7]
Case Duration, Mean [Min- Max] (Days)	3,922 [432 – 5,682]	3,821 [378 – 5,684]
Case Duration, Median [IQR] (Days)	4,134 [3,074 – 4,949]	3,999 [2,893 – 4,954]

The two-trace variant are being compared using Process Comparator plug-in ProM

I) Process Metrics: Frequency



II) Process Metrics: Elapsed time



The descriptive statistics of the compared logs:

State Label	Variant I	Variant II
	Average Elapsed time (days)	
"[Polypharmacy], [Mild]"	**992**	415
"[Hypertension], [Mild]"	444	**723**
"[Mild], [Fall]"	**1,695**	1,420
"[Fall], [Moderate]"	**2,738**	2,339
"[Mild], [Moderate]"	**1,335**	1,101
"[Moderate], []"	**3,395**	3,083
"[Moderate], [Fall]"	**2,803**	2,421
"[Moderate], [Severe]"	**2,403**	2,009
"[Severe], []"	**3,777**	3,506

Conclusion

The analysis of frailty progression in the association with deficits of concern (fall, hypertension and polypharmacy) reveals several important information. Further discussion on the finding and their evaluation has been discussed ion Section 6.4.4.

A.7 Experiment 7: Bradford Dataset (Confirmatory Analysis)

UNIVERSITY OF LEEDS School of Computing	**EXPERIMENT DOCUMENTATION** **Experiment title:** Comparative Analysis (as part of confirmatory analysis) between proposed cut-off points and cut-off points used in literature **Researcher's name:** Nik Farid	**Date of experiment** 17 February 2021 **Experiment code** BD_CA_C7E7

Area of investigation

This experiment is determining the association of focused frailty deficits within frailty stages with the proposed cut-off points. This experiment aims in determining the difference between two sets of cut-off points to define the frailty categories. The aim of the experiments to determine the difference of frailty progression within frailty stages between two sets of cut-off points.

Data source

The Bradford dataset

Research question

1) What are the differences in frailty progression and its association with focused deficits between cut-off points used in literature and the proposed cut-off points within frailty stages?

Hypothesis

The discovered frailty progression model and its association with fall, hypertension and polypharmacy can be used as initial step in analysing the relatedness with frailty severity utilising process mining approach.

Method:

The general method used is the Process Mining Project Methodology (PM2) which includes Stage I: Planning, Stage II: Extraction, Stage III: Data Transformation and Loading and Stage IV: Mining and Analysis and Stage V: Evaluation.

I: Planning - Three research questions are addressed (as above) to investigate the frailty progression within deficits of concern. The questions are based on the aim to determine the association of deficits of concern within frailty stages.

II: Extraction – Minimum aged of 65 years elderly patient at the start of study duration within year 2003 – 2018. Three additional exclusions are:

iv. Patient has at least one year record within study duration
v. Patient aged over 85 years old who their final records in either fit or mild category
vi. Maximum average accumulation of frailty deficits is only three

A total of 12,395 patients with three deficits of concern (fall, hypertension, and polypharmacy) is 8,547 and without is 3,848.

III: Data Transformation and Loading – Creating view based on the structure of the data and research aim. Additional transformation steps are as follows:

iv. Identification of frailty score – the first occurrence of each deficit associated with frailty was identified at each visit to GP practice
v. Log enriching – additional events are extended into the initial event log which are frailty stages; fit, mild, moderate, and severe.
vi. Securing the events sequence – the order of the events must always start with frailty stage event, if frailty stage event and frailty deficits events shared similar timestamp.

Once events logs are created for sub-cohort with deficits of concern and without, save them in .csv file. Then load it into Disco and extract .XES files to load into ProM tool.

Results

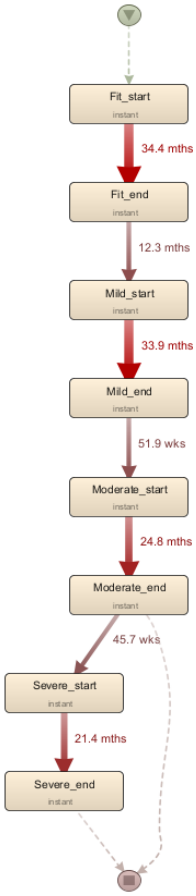
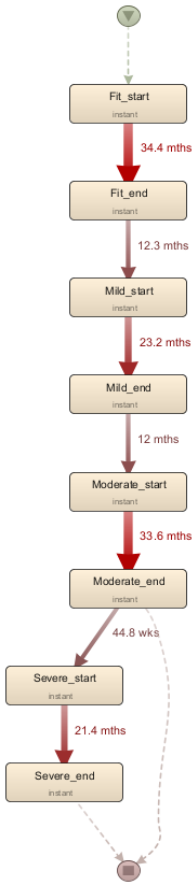
Stage IV: Mining and Analysis

Result and discussion

3. The two sub-cohorts of with and without deficits of concerns are being analysed.
   A) Descriptive statistics of two sub-cohorts:

	Clegg's Cut-off Points	**Proposed Cut-off points**
# Patients	8,547	8,547
# Events	30,754	30,754
Events per patient	3.6 (~4)	3.6 (~4)
Trace Fitness	0.96	0.96
Precision	1.00	1.00
Generalisation	0.77	0.77
Frailty Stages	Median Duration (IQR) in years	
Fit	2.2 (1.1 – 3.9) N = 8,547	**2.2 (1.1 - 3.9)** N = 8,547
Mild	**2.3 (1.2 - 4.0)** N = 8,514	1.4 (0.6 – 2.7) N = 8,514
Moderate	**1.6 (0.5 - 3.1)** N = 7,023	2.3 (1.0 –4.2) N = 7,025
Severe	0.9 (0.0 – 2.8) N = 3,335	0.9 (0.0 – 2.8) N = 3,335
Transition Point	Median Duration (IQR) in years	
1	0.6 (0.2 – 1.4) N = 8,514	**0.6 (0.2 – 1.4)** **N = 8,514**
2	0.6 (0.2 – 1.4) N = 7,023	0.5 (0.2 – 1.4) N = 7,025
3	0.5 (0.2 – 1.2) N = 3,335	0.5 (0.1 – 1.2) N = 3,335

B) The process models with performance view of two cut-off points:

Clegg's Cut-off Points	Proposed Cut-off Points
	

C) The Descriptive statistics between two sets of cut-off points within different sub-cohorts based on age classification

	Clegg's Cut-off Points			Proposed Cut-off Points		
Age-groups	(65 – 74)	(75 – 84)	(85++)	(65 – 74)	(75 – 84)	(85++)
# Patients	3,395	4,289	863	3,395	4,289	863
# Events	39,757	50,426	9,926	39,757	50,426	9,926
Events per patients	11.7 (~12)	11.8 (~12)	11.5 (~12)	11.7 (~12)	11.8 (~12)	11.5 (~12)
	Duration of Frailty Stages Median (IQR) (in years)					
Fit	2.7 (1.3 - 4.6) N = 3,395	2.1 (1.0 - 3.7) N = 4,289	1.8 (0.9 - 3.0) N = 863	2.7 (1.3 - 4.6) N = 3,395	2.1 (1.0 - 3.7) N = 4,289	1.8 (0.9 - 3.0) N = 863
Mild	2.7 (1.4 – 4.5) N = 3,373	2.2 (1.1 - 3.8) N = 4,289	1.6 (0.9 - 2.8) N = 863	1.7 (0.8 - 3.1) N = 3,373	1.3 (0.6 - 2.6) N = 4,278	1.0 (0.5 - 1.9) N = 863
Moderate	1.8 (0.6 – 3.4) N = 2,727	1.6 (0.5 - 3.0) N = 3,519	1.1 (0.2 - 2.2) N = 777	2.6 (1.1 – 4.7) N = 3,007	2.2 (0.9 – 4.0) N = 3,847	1.8 (0.8 – 3.1) N = 806

Severe	1.1 (0.0 – 3.0) N = 1,354	0.9 (0.0 - 2.8) N = 1,699	0.4 (0.0 - 1.8) N = 282	1.1 (0.0 - 3.0) N = 1,354	0.9 (0.0 - 2.8) N = 1,699	0.5 (0.0 - 1.8) N = 282
	Duration of Transition Points Median (IQR) (in years)					
1	0.7 (0.2 - 1.6) N = 3,373	0.6 (0.2 - 1.3) N = 4,278	0.4 (0.1 - 1.0) N = 863	0.7 (0.2 - 1.6) N = 3,373	0.6 (0.2 - 1.3) N = 4,278	0.4 (0.1 - 1.0) N = 863
2	0.7 (0.2 – 1.6) N = 3,373	0.5 (0.2 - 1.3) N = 3,519	0.4 (0.1 – 1.0) N = 777	0.7 (0.2 - 1.6) N = 3,007	0.6 (0.2 - 1.3) N = 3,847	0.4 (0.1 – 1.0) N = 806
3	0.5 (0.2 - 1.3) N = 1,354	0.5 (0.2 - 1.2) N = 1,699	0.4 (0.1 - 1.1) N = 282	0.5 (0.2 - 1.3) N = 1,354	0.5 (0.1 - 1.2) N = 1,699	0.4 (0.1 - 1.1) N = 282

D) Median (Mean) duration (in years) between each frailty stages and the transition between stages

	Clegg's Cut-off Points			Proposed Cut-off Points		
	65 – 74	75 – 84	85++	75 – 84	65 – 74	85++
Fit (Start) - Fit (End)	2.7 (3.3)	2.1 (2.7)	1.8 (2.2)	2.7 (3.3)	2.1 (2.7)	1.8 (2.2)
Fit (End) - Mild (Start) [TP1]	0.7 (1.2)	0.6(4.1)	0.4 (3.3)	0.7 (1.2)	0.6 (4.1)	0.4 (3.3)
Mild (Start) - Mild (End)	2.7 (3.2)	2.2 (2.7)	1.6 (2.1)	1.7 (2.2)	1.4 (1.8)	1.0 (1.4)
Mild (End) - Moderate (Start) [TP 2]	0.7 (1.1)	0.5 (0.9)	0.4 (0.8)	0.7 (1.1)	0.6 (1.0)	0.4 (0.7)
Moderate (Start) - Moderate (End)	1.8 (2.3)	1.6 (2.0)	1.1 (1.5)	2.7 (3.1)	2.2 (2.7)	1.8 (2.2)
Moderate (End) - Severe (Start) [TP 3]	0.5 (0.9)	0.5 (0.9)	0.4 (0.8)	0.5 (0.9)	0.5 (0.8)	0.4 (0.8)
Severe (Start) - Severe (End)	1.1 (1.9)	0.9 (1.8)	0.4 (1.2)	1.1 (1.9)	0.9 (1.8)	0.4 (1.2)

E) The cut-off points of accumulated deficits

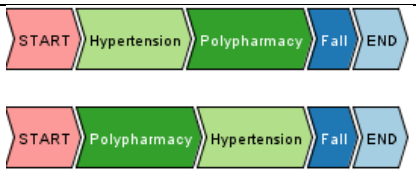
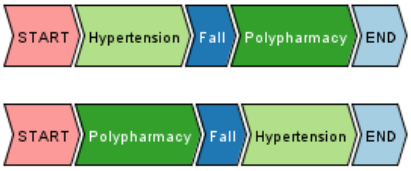
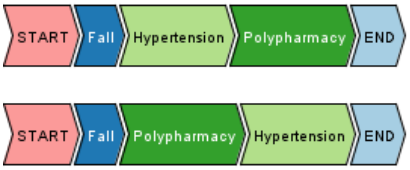
Frailty Stages	Range of Accumulated Deficits			
	Clegg's	Count of Deficits	Proposed	Count of Deficits
Fit	0 – 4	4	0 – 4	4
Mild	5 – 8	4	5 – 7	3
Moderate	9 – 12	4	8 – 13	6
Severe	13 – 36	24	14 – 36	23

Observation: Frailty stages with different number of count of deficits from Clegg's produced different results from the proposed cut-off points. The differences can be observed between the start of mild stages until the end of moderate stage. (The highlighted stages from the above Table D and E)
Finding: Stages with high number of count of deficits showed higher median duration at all sub-cohorts (of age classification). Pattern of frailty progression is decreasing from fit until severe stages using the Clegg's cut-off points at all sub-cohorts. While the duration of proposed cut-off points is

decreasing from fit until mild, it is experiencing slightly increasing duration from mild to moderate stage (range between 0.8 -1.0) at all sub-cohorts.

F) The Pattern of Sequence (PoS) of age classification sub-cohorts from both sets of cut-off points are examined to determine the association with deficits of concern e.g., fall, hypertension and polypharmacy within frailty stages. The PoS are based on the previous experiment (Section 6.4) to investigate the association of deficits of concern.

G) Pattern of Sequence (PoS)

Pattern of Sequence	Trace Variants
I	
II	
III	

H) Descriptive statistics of pattern of sequence I, II and III for each sub-cohort.

Characteristics	LR	Proposed	LR	Proposed	LR	Proposed
	65 – 74		75 – 84		85++	
	PoS I					
# Patient (%)	2,266 (26.5%)		2,566 (30.0%)		424 (5.0%)	
# Events (%)	14,198	14,115	16,061	15,980	2,678	2,629
# Trace Variant	61	61	64	61	48	42
Case Duration, Mean [Min- Max] (Years)	2.0 [0.0 – 14.1]	2.0 [0.0 – 14.1]	1.7 [0.0 – 13.0]	1.7 [0.0 – 13.0]	1.3 [0.0 – 11.9]	1.3 [0.0 – 11.9]
Case Duration, Median [IQR] (Years)	1.2 [0.1 – 3.1]	1.1 [0.1 – 3.0]	1.0 [0.1 – 2.5]	1.0 [0.1 – 2.5]	0.7 [0.0 – 2.0]	0.6 [0.0 – 1.9]
	PoS II					
# Patient (%)	762 (8.9%)		1,191 (13.9%)		291 (3.4%)	
# Events (%)	4,636	4,646	7,319	7,319	1,794	1,776
# Trace Variant	56	52	57	54	46	45
Case Duration, Mean [Min- Max] (Years)	1.9 [0.0 – 14.4]	1.8 [0.0 – 14.4]	1.6 [0.0 – 14.5]	1.6 [0.0 – 14.4]	1.3 [0.0 – 12.1]	1.2 [0.0 – 11.8]
Case Duration, Median [IQR] (Years)	1.0 [0.1 – 2.9]	1.0 [0.1 – 2.8]	0.8 [0.1 – 2.4]	0.8 [0.1 – 2.3]	0.7 [0.1 – 2.0]	0.7 [0.0 – 1.9]
	PoS III					
# Patient (%)	367		532		148	

	(4.3%)		(6.2%)		(1.7%)	
# Events (%)	2,200	2,205	3,272	3,254	912	899
# Trace Variant	44	43	52	51	32	31
Case Duration, Mean [Min- Max] (Years)	1.9 [0.0 – 12.9]	1.9 [0.0 – 12.9]	1.5 [0.0 – 11.2]	1.4 [0.0 – 10.7]	1.2 [0.0 – 10.1]	1.1 [0.0 – 10.1]
Case Duration, Median [IQR] (Years)	1.2 [0.2 – 3.0]	1.1 [0.1 – 2.9]	0.8 [0.1 – 2.2]	0.8 [0.1 – 2.1]	0.7 [0.0 – 1.6]	0.6 [0.0 – 1.6]

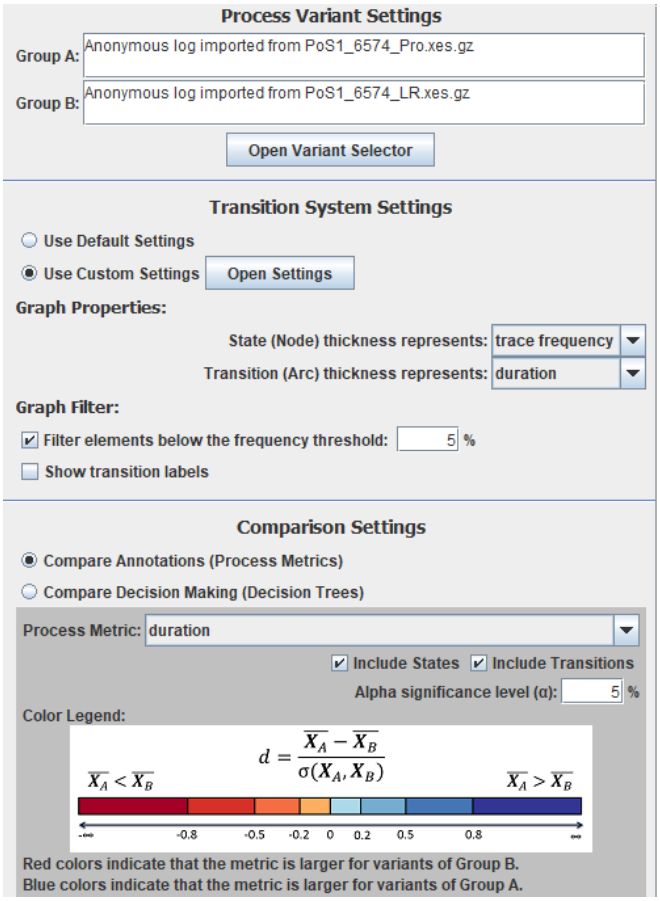
** The highlighted cells present the dominant PoS among the frail elderly.

I) The number of events is difference despite the same number of patients in each sub-cohort, is because the proposed cut-off points identify the minimum accumulated deficits for severe is 14. Meanwhile in Clegg's cut-off points identify 13 as severe. In the dataset, while applying Clegg's cut-off points, we identify higher number of severe category patient as compared to when applying the proposed cut off points. (Means: the dataset with high severe category patient with minimum deficits of 13.

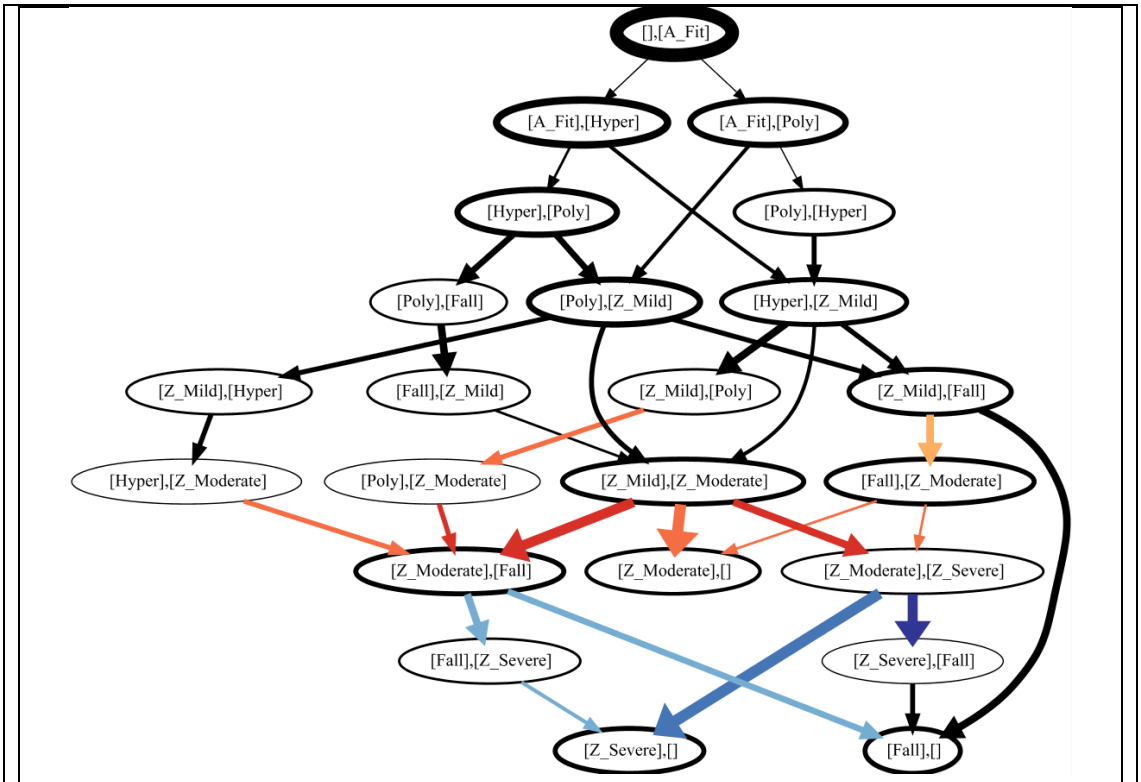
J) The dominant PoS are being compared as it has different variants with the same process. We use Process Comparator Plug-in in ProM. Statistically significance of states label interval time was measured and presented in Table below.

Process Variant Graph Visualization

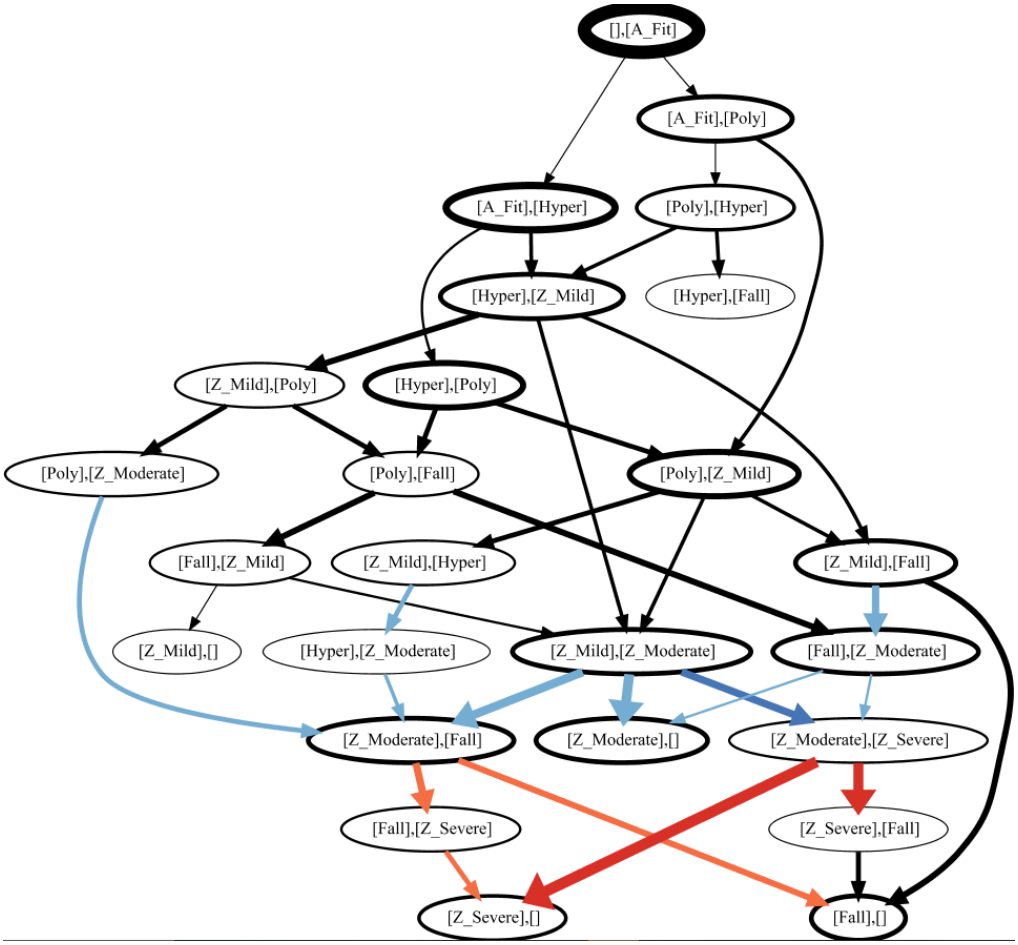
Process Variant Settings



65 – 74 (Clegg's VS Proposed)



75 – 84 (Clegg's VS Proposed)



The descriptive statistics of the compared logs:

i) Logs from the Clegg's and Proposed sub-cohorts of 6475. The numerical information in the table describes the % of Trace Frequency (Duration of Interval in years). The states label associated with the deficits are concern are listed in the table for the two dominant sub-cohorts:

State Label	Clegg's	Proposed	State Label	Clegg's	Proposed
(65 – 74)			(75 – 84)		
[Mild, Fall] > [Moderate]	**27.1 (3.4)**	24.6 (2.9)	*[Mild, Fall] > [Moderate]*	**26.5 (2.9)**	23.7 (2.3)
[Mild, Moderate] > [Fall]	16.8 (4.6)	**25.5 (3.4)**	*[Mild, Moderate] > [Fall]*	15.6 (3.8)	**25.6 (3.0)**
[Moderate] > [Fall]	11.4 (2.0)	**24.8 (2.8)**	*[Moderate] > [Fall]*	11.9 (1.9)	**25.3 (2.5)**
[Fall] > [Moderate]	23.3 (1.2)	**24.2 (0.9)**	*[Fall] > [Moderate]*	**24.8 (1.1)**	24.1 (0.7)
[Fall] > [Severe]	**19.9 (1.2)**	19.4 (1.8)	*[Fall] > [Severe]*	**19.5 (1.2)**	19.1 (1.9)
[Moderate, Fall] > [Severe]	**18.5 (2.7)**	16.6 (3.6)	*[Moderate, Fall] > [Severe]*	**18.2 (2.3)**	16.6 (3.1)
[Fall, Moderate] > [Severe]	**11.7 (1.1)**	5.5 (0.8)	*[Fall, Moderate] > [Severe]*	**11.4 (0.9)**	5.7 (0.5)

Conclusion

The analysis of frailty progression in the association with deficits of concern (fall, hypertension and polypharmacy) reveals several important information. Further discussion on the finding and their evaluation has been discussed ion Section 6.4.4.