Exploring the use of routine healthcare data through process mining to inform the management of musculoskeletal diseases

Samantha Jane Sykes

Submitted in accordance with the requirements for the degree of

Doctor of Philosophy

The University of Leeds

School of Medicine

November 2022

Intellectual Property and Publication Statement

The candidate confirms that the work submitted is her own and that appropriate credit has been given where reference has been made to the work of others.

This copy has been supplied on the understanding that it is copyright material and that no quotation from the thesis may be published without proper acknowledgement.

The right of Samantha Jane Sykes to be identified as Author of this work has been asserted by Samantha Jane Sykes in accordance with the Copyright, Designs and Patents Act 1988.

Acknowledgements

My heartfelt thanks to my five supervisors, Professor Philip Conaghan, Dr Sarah Kingsbury, Mr Owen Johnson, Dr Mar Pujades Rodriguez and Professor Paul Baxter, for their guidance, encouragement and continued faith in me over these last four years. Without them this thesis would not have been possible, they have made my experience pleasurable and inspiring.

Research in this thesis was carried out with the support of a post-graduate process mining research team. This team was led by Mr Owen Johnson and included Amirah Alharbi, Angelina Kurniati, Frank Fox, Guntur Kusuma, Nik Fatinah and Samantha Crossfield. A special thank you must go to Dr Mark Dixon, Dr Thamer Ba Dhafari and Dr Eric Rojas who have devoted their knowledge and precious time.

I would like to thank all members of the technical support team within the School of Computing and the School of Medicine at the University of Leeds and the SAIL databank at Swansea University. Without them, my work would have come to a standstill on many occasions. I cannot forget Judi, who has since retired, and Gaynor in the School of Computing admin team, as nothing was ever too much trouble and they both brought fun and laughter to every day. I would like to thank everybody at ADI, especially John Eaglesham and Simon Bramwell for welcoming me into their team and for giving me such a fantastic opportunity.

I owe a huge thank you to my family: my son Oliver, who has grown into an amazing young man and has made me so enormously proud; to my parents Lynne and Dennis, I couldn’t wish for better; and to my aunt Julie for all their continued support and belief. And finally to Mark, it would not have been possible to undertake this amount of work had it not been for his constant support and copious supply of food.

Abstract

Healthcare informatics can help address some of the challenges faced by both healthcare providers and patients. The medical domain is characterised by inherently complex and intricate issues, data can often be of poor quality and novel techniques are required. Process mining is a discipline that uses techniques to extract insights from event data, generated during the execution of processes. It has had good results in various branches of medical science but applications to musculoskeletal diseases remain largely unexplored.

This research commenced with a review of the healthcare and technical literature and applied a variety of process mining techniques in order to investigate approaches to the healthcare plans of patients with musculoskeletal conditions. The analysis involved three datasets from: 1) a private hospital in Boston, US, where data was used to create disease trajectory models. Results suggest the method may be of interest to healthcare researchers, as it enables a more rapid modelling and visualisation; 2) a mobile healthcare application for patients receiving physiotherapy in Sheffield, UK, where data was used to identify possible indicators for health outcomes. After evaluation of the results, it was found that the indicators identified may be down to chance; and 3) the population of Wales to explore knee pain surgery pathways. Results suggest that process mining is an effective technique.

This work demonstrates how routine healthcare data can be analysed using process mining techniques to provide insights that may benefit patients suffering with musculoskeletal conditions. This thesis explores how strict criteria for analysis can be performed. The work is intended to expand the breadth of process mining methods available to the data science community and has contributed by making recommendations for service utilisation within physiotherapy at Sheffield Hospital and helped to define a roadmap for a leading healthcare software company.

Table of Contents

List of Tables

List of Figures

List of Terminology

Term	Definition
activity	a well-defined step or task
artefact	any tangible by-product produced during mining and analysis, e.g. a model or a chart
behaviour	the pattern of events created by cases in the event log over time
care pathway	the flow between activities applied to a group of patients with the same condition
case	an individual process instance
conformance checking	a type of process mining that is used to compare the 'real life' information in an event log to a process model and vice versa
control flow	a perspective concerned with the ordering of activities
enhancement	a type of process mining that is used to extend or improve an existing process model
event	an activity within an EHR or an instance of an event type
event log	a collection of timestamped events relating to a case
event type	a classification identifying one of various types of event
fitness	the ability of a process model to accurately reproduce the behaviour seen in the event log
generalisation	the ability of a process model to allow for future behaviour not seen in the event log
lateral	toward the side or away from the centre
medial	toward the inner or central
PhysioWorks	community based physiotherapy service
precision	a model should not allow for behaviour unrelated to that in the event log
process	a set of activities taken to achieve a particular end
process discovery	a type of process mining that is used to produce a process model using only an event log
process mining	a method for data analytics that can offer insights to help understand and improve complex processes
process model	an abstraction of a process which reflects pathways
proximal	near to the point of attachment or reference point
sidedness	reference to the left or right side of the body
trace	the sequence of activities for a case as recorded in the event log
variant	a unique pathway or trace

List of Abbreviations

Abbreviation	Term
ACR	American College of Rheumatology
ADI	Advanced Digital Innovation
BIDMC	Beth Israel Deaconess Medical Center
BPMN	Business Process Modelling and Notation
CVD	Cardiovascular Disease
CITI	Collaborative Institutional Training Initiative
DBMS	Database Management System
DNA	Did Not Attend
GBD	Global Burden of Disease
EHR	Electronic Health Record
ERP	Enhanced Role Physiotherapists
ETL	Extract, Transform and Load
EULAR	European League Against Rheumatism
ICD	International Classification of Diseases
IDE	Integrated Development Environment
KPI	Key Performance Indicator
LBP	Lower Back Pain
LIRMM	Leeds Institute of Musculoskeletal Medicine
LTC	Long Term Condition
MIMIC-III	Medical Information Mart for Intensive Care III
MRI	Magnetic Resonance Imaging
MSK	Musculoskeletal
MPE	Multi-Perspective Process Explorer
NHS	National Health Service
NICE	National Institute for Health and Care Excellence
NSAID	Non-Steroidal Anti-Inflammatory Drug
OA	Osteoarthritis
OMG	Object Management Group
OPCS	Classification of Interventions and Procedures
PEDW	Patient Episode Database for Wales
PNML	Petri Net Markup Language
PCHC	Paretian Classification of Health Change
PCP	Primary Care Practitioner
PM^2	Process Mining Project Methodology
PROM	Patient Reported Outcome Measure
QOF	Quality and Outcomes Framework
RDBMS	Relational Database Management System
SAIL	Secure Anonymised Information Linkage
SQL	Structured Query Language

STHT	Sheffield Teaching Hospitals NHS Foundation Trust
STROBE	Strengthening the Reporting of Observational Studies in Epidemiology
TKA	Total Knee Arthroplasty
TKR	Total Knee Replacement
ULT	Urate-Lowering Therapy
UML	Unified Modelling Language
VAS	Visual Analogue Scale
WDS	Welsh Demographics Service
WHO	World Health Organisation
WLGP	The Welsh Longitudinal General Practice dataset

Chapter 1
Introduction

1.1 Overview

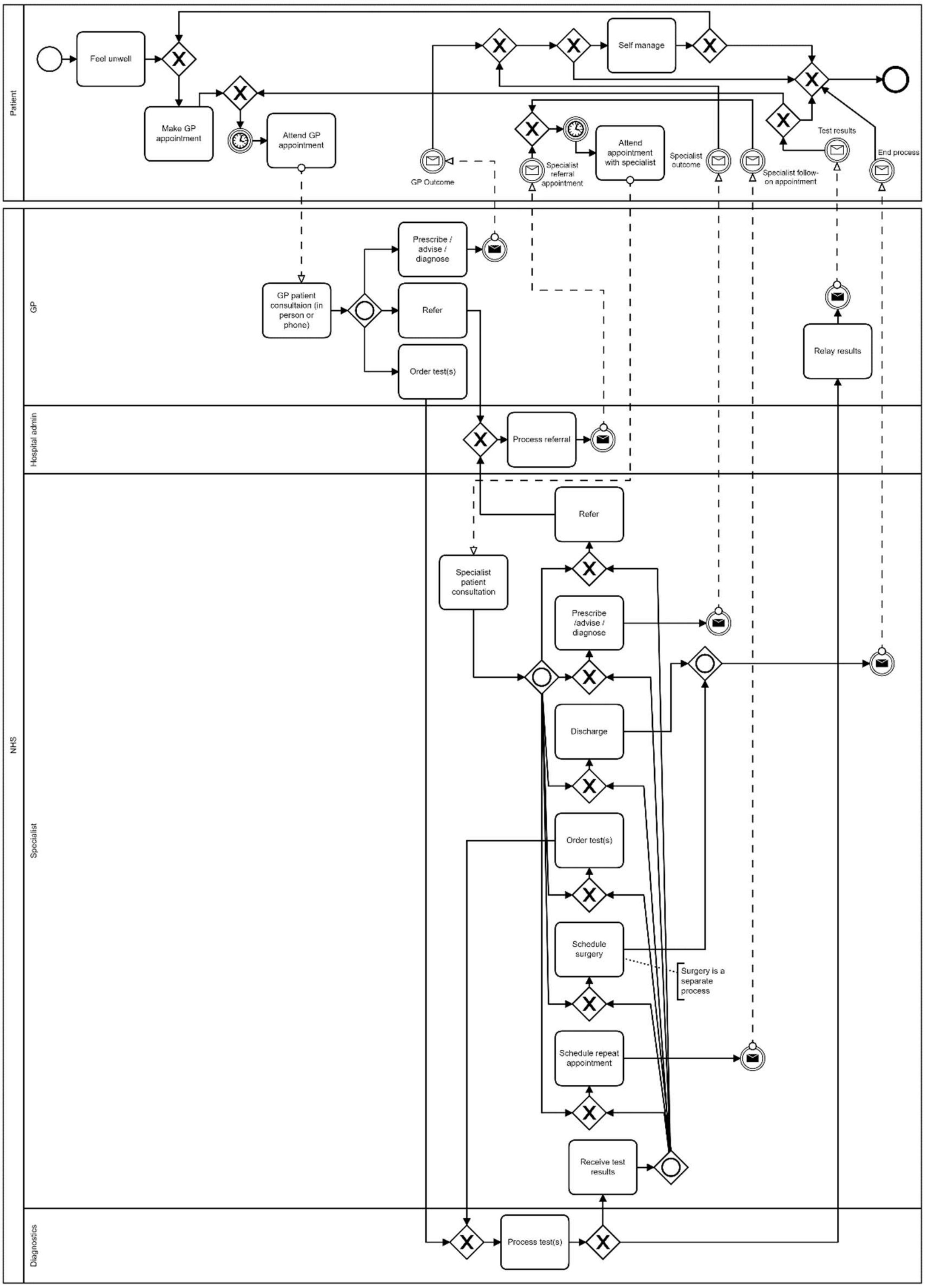
Long-term conditions (LTC), also known as chronic diseases, are conditions for which there is no cure, meaning they must be managed with drugs or by other treatment methods. The global prevalence of patients with LTCs is rising due to increasingly obese and ageing populations [1]–[3]. Some examples of LTCs include arthritis, diabetes, chronic obstructive pulmonary disease and hypertension. Musculoskeletal (MSK) conditions, such as arthritis, make up a large percentage of all LTCs. They often cause pain and limit mobility by affecting the muscles, bones, joints and soft tissue [4] in addition to significantly contributing to patient disability [5]. Most MSK cases are dealt with in primary care (see Section 2.1) and account for an estimated 30 percent of all general practitioner (GP) consultations in England [6].

Due to the rise in LTCs, new methods for managing and delivering health care are desperately needed to help minimise the effects on health and social care services [7] and to improve the quality of life for people suffering with these conditions. Most modern health care is facilitated by the use of computerised information systems, where routine patient data is stored in the form of structured electronic health records (EHR). EHRs are a cost-effective source of longitudinal patient data that often includes information such as patient demographics, referrals, laboratory test results, diagnoses, medications, procedures and billing details. A better understanding of different healthcare processes through the analysis of EHR data could lead to improvements in patient health management [8], [9] and in turn, the quality of life for people living with MSK conditions. However, a major challenge when working with healthcare processes is that they are extremely complex, ad-hoc, dynamic and multidisciplinary in nature, making them difficult to analyse using existing techniques [10]. In addition, as more healthcare providers migrate from paper based systems to

EHRs and data volumes grow, often with low levels of quality, existing analysis methods need to improve.

Process mining is an emerging method for data analytics that can offer novel insights to help understand and improve healthcare processes [11], [12]. It can be used for large volumes of complex data in order to help discover, monitor, and improve processes by analysing data from EHRs in the form of event logs [12]. An event log is a collection of timestamped events relating to a case, such as a patient, where an event is an activity within an EHR, such as a referral into secondary care [12]. Process mining techniques are often used to discover process models, which are step-by-step representations of a process from a certain perspective in an abstract form [13]. Within these models processes are constructed from a sequence of connected activities and individual cases from the event log. These individual cases exhibit distinct *behaviour* as they take different routes through the model. In a healthcare setting, the flow between activities applied to a group of patients with the same condition is known as a care pathway [14]. Using process mining techniques, care pathways can be established using ‘real-life’ patient data stored in EHRs. These care pathways can help healthcare professionals to better understand possible indicators of patient health outcomes, in turn helping to provide insights that may benefit patients suffering with musculoskeletal conditions. Figure 1.1 presents an example of a care pathway created using the standard Business Process Modelling and Notation (BPMN) [15] (Section 3.3.2).

Figure 1.1 Representation of a care pathway



The care pathway presented in Figure 1.1 presents the logical flow of activities between a patient and the National Health Service (NHS) in the United Kingdom (UK). The process begins with a patient making a GP appointment and ends when no more can be done for the patient. The following section provides an illustrated example by introducing Jack. Jack's journey can be navigated using the process model in Figure 1.1.

1.2 An illustrative example of the care pathway

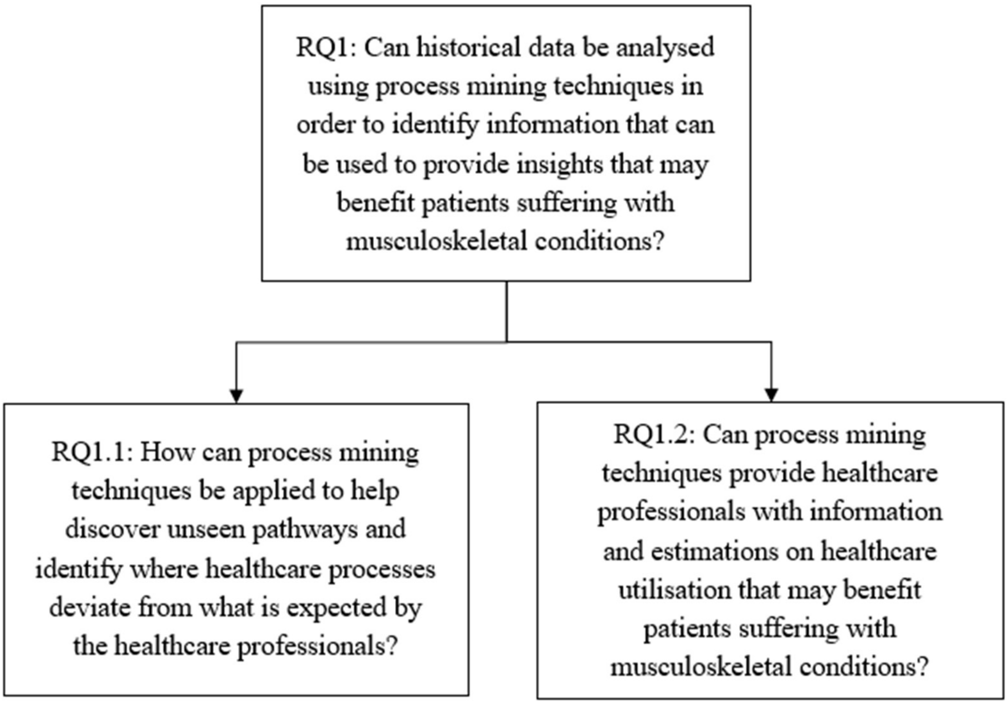
An illustrative example is provided by following the journey of Jack, a 50 year old male office manager. Jack had been suffering from neck pain for two years, though recently the pain had started to radiate across his shoulder and down his right arm, causing tingling and numbness. Jack made and appointment with his GP. After a review of Jack's history and symptoms, the GP performed a physical and neurological examination before ordering an x-ray, referring him to therapy services and prescribing Jack ibuprofen. Following the x-ray, Jack's GP informed him that the images showed a bulging C5/6 disc. One week later Jack attended his appointment at therapy services. After a neurological examination, Jack was given some neck strengthening exercises and a follow-up appointment in four weeks' time. At this appointment Jack explained how his symptoms had worsened, leaving him unable to work. Jack was urgently sent for a magnetic resonance imaging (MRI) scan. On receipt of the results, which confirmed the initial diagnosis of cervical spondylosis myelopathy (narrowing of the spinal canal with spinal cord compression), therapy services referred Jack to a neurosurgeon to discuss potential surgery. After speaking with the neurosurgeon surgery was scheduled, as both Jack and the surgeon agreed that it was the best option.

1.3 Hypothesis, research question and objectives

The aim of this research is to explore the application of process mining to routine healthcare data in order to provide insights that may benefit patients suffering with musculoskeletal conditions. To exploit the benefits and limitations of process mining, different techniques will be applied across a diverse range of routinely collected datasets.

Toward this aim, the key hypothesis is that *routine healthcare data can be analysed using process mining techniques in order to identify information that can be used to provide insights that may benefit patients suffering with musculoskeletal conditions*. The primary research questions are shown in Figure 1.2 and were developed by breaking down this hypothesis.

Figure 1. 2 Primary research questions



RQ1: Can historical data be analysed using process mining techniques in order to identify information that can be used to provide insights that may benefit patients suffering with musculoskeletal conditions?

RQ1.1: How can process mining techniques be applied to help discover unseen pathways and identify where healthcare processes deviate from what is expected by the healthcare professionals?

RQ1.2: Can process mining techniques provide healthcare professionals with information and estimations on healthcare utilisation that may benefit patients suffering with musculoskeletal conditions?

These primary research questions will be further broken for each study in chapters 5, 6 and 7. The high-level objectives set to realise the aim and answer these research questions are:

RO1. Identify appropriate datasets.

RO2. Identify appropriate automated and process mining techniques.

RO3. Apply the process mining techniques to the datasets.

RO4. Analyse the results.

RO5. Discuss and summarise whether the results satisfy the hypothesis.

Contribution for this research programme is towards both the process mining in healthcare community and wider health informatics research, focusing primarily on MSK conditions.

1.4 Study approach

The approach taken with this research was to carry out three individual studies in order to test the hypothesis using an adaptation of the Process Mining Project Methodology (PM^2) [16]. The main inputs consisted of three different datasets, one for each study. Other inputs included documentation for the datasets, existing process models and knowledge gained through the literature and via discussions with various domain experts. Process mining techniques were used in the first study to create disease trajectories, in the second study to discover patterns of care between patients with different outcome measures and in the third study to create a tool for the analysis of knee pain surgery. Clinical and technical evaluation, along with the comparison of results with those from similar studies was carried out.

1.4.1 Data sources

The datasets used in this research programme originate from three completely different data sources. These data sources were: the Medical Information Mart for Intensive Care III (MIMIC-III) research database [17], [18] (USA); the MyPathway mobile healthcare application [19] (England); and the Secure Anonymised Information Linkage (SAIL) databank [20] (Wales). These datasets provided the author with the opportunity to experience different opportunities and challenges when applying process mining techniques to the data.

The first data source was the MIMIC-III database. MIMIC-III comprised of over forty thousand de-identified EHRs for patients who attended the critical care units of the Beth Israel Deaconess Medical Center (BIDMC) [21] in Boston, USA between 2001 and 2012. The BIDMC is a private teaching hospital for the Harvard Medical School, formed from a merger in 1996 between Beth Israel Hospital and New England Deaconess Hospital. The center is part of Beth Israel Lahey Health, a healthcare system for medical centers, teaching hospitals and community and speciality hospitals. The city of Boston covers an area of 48 square miles and had an estimated population of 710,195 in 2020 [22]. No other research has been published using MIMIC-III data for process mining MSK data, though 22 articles have been published using MIMIC-III data for process mining. Kurniati et al. used MIMIC-III data to provide an assessment of data quality issues for process mining in healthcare [23] and also for a study using process mining techniques in oncology [24]. Alharbi et al. used patients with diabetes from MIMIC-III to test a new method for variation reduction in clinical pathways data [25]. Data from dental patients in MIMIC-III was used by Fox et al. when developing a care pathway data quality framework for process mining [26]. Using the MIMIC-III data provided many opportunities, these included: 1) access to data from a privately funded hospital in the USA; 2) working with medical data coded in ICD-9 format (see Section 2.2); 3) continuous access to the entire database which allowed for an iterative development approach; and 4) future research is more reproducible, as all data is freely available.

The second data source was from the mobile health application, MyPathway. The privately owned application is used in various hospital settings across the UK and Spain, though for this study it is used by patients attending the Sheffield Teaching Hospitals NHS Foundation Trust (STHT), PhysioWorks service in Sheffield, England. STHT manages five NHS adult hospitals in Sheffield [27] and hosts a number of community services including the MSK service, PhysioWorks [28]. From April 2017 all patients attending PhysioWorks were given the opportunity to use the MyPathway application to help manage their condition. Sheffield is a large UK city covering an area of 142 square miles, with a largely white population (84%) of approximately 730,000 in 2020 [29]. This thesis is the first publication using MyPathway data. Using the MyPathway dataset provided many unique opportunities and challenges, these included: 1) working within a small, cutting-edge software development company specialising in healthcare, with resources available to support innovation in the field of data science and process mining; 2) accessing data collected from an MSK mobile application; 3) accessing Patient Reported Outcome Measures (PROM) data (see Section 2.3.2); 4) exploring data never previously used within a research setting; 5) experiencing some of the difficulties related to process mining data originating from a non-standard database; 6) unlike the MIMIC-III and SAIL datasets, using raw and un-curated data, locked inside NoSQL files and the minds of the development team; and 7) experiencing some of the difficulties associated with navigating a real-world commercial situation, with competing demands from an NHS healthcare provider.

The final data source was the SAIL databank. SAIL is a data linkage research platform for population health and social care data, established in 2007. It contains de-identified linked primary and secondary care EHR data for patients from 1991 to the current day. Seventy-seven percent of GP practices and all of NHS hospitals throughout Wales submit their data to SAIL [30]. Wales is a country in the UK mainland and covers an area of 8,006 square miles, with a predominantly white population of approximately 3,152,879 in 2019 [31]. This thesis is the first publication on process mining using data from the SAIL Databank. Using the SAIL dataset provided many opportunities, these included: 1) using linked primary and secondary care data, which may be important when working with diagnosis and surgery data; 2) working with data never previously published in a process mining study 3) having access to a full support team where technical and medical issues could be discussed; and 4) to work with a large dataset of over two million patients, covering an entire country.

1.4.2 Overview of the datasets

This section provides an introduction to each of the datasets used during the three studies. A more detailed description is available in Section 4.2. Table 1.1 presents the total number of patients in each dataset, along with the number extracted for analysis in each study.

Table 1.1 Datasets used in this research

Dataset	Study
MIMIC-III (n=46,520)	Chapter 5, cardiovascular disease trajectory (n=32,457)
MyPathway (n=119,266)	Chapter 6, knee pain patients (n=436), spinal pain patients (n=721)
SAIL	Chapter 7, knee surgery pathways (n=11,289)

(2,035,913)	

**n = number of patients*

1) The MIMIC-III dataset

The entire MIMIC-III database was reconstructed onto a local machine. Information included patient demographics, vital sign measurements, procedures, medications, caregiver notes, test results, imaging reports, and mortality between 2001 and 2012. MIMIC-III is made available by the MIT Laboratory for Computational Physiology [32]. Access to this freely accessible database was granted through an online research ethics approval process. The author was required to complete the Collaborative Institutional Training Initiative (CITI) 'Data or Specimens Only Research' course [33]. The course contained nine modules centring around ethical and professional conduct when performing research involving human subjects.

2) The MyPathway dataset

The MyPathway dataset was provided by developers of the application, Advanced Digital Innovation (ADI) UK Limited [34], in the form of text files. The files consisted of anonymised time stamped, patient-level healthcare management events recorded between 15/05/2017 and 12/08/2019. ADI are a privately owned software company, providing software solutions to healthcare providers and are located in Saltaire, UK. Access was granted to the researcher upon completion of the NHS Data Security Awareness Level 1 course [35]. Ethical approval was granted (MREC17-108) for this study via the Medicine and Health University Ethics Review team on 14/02/2020.

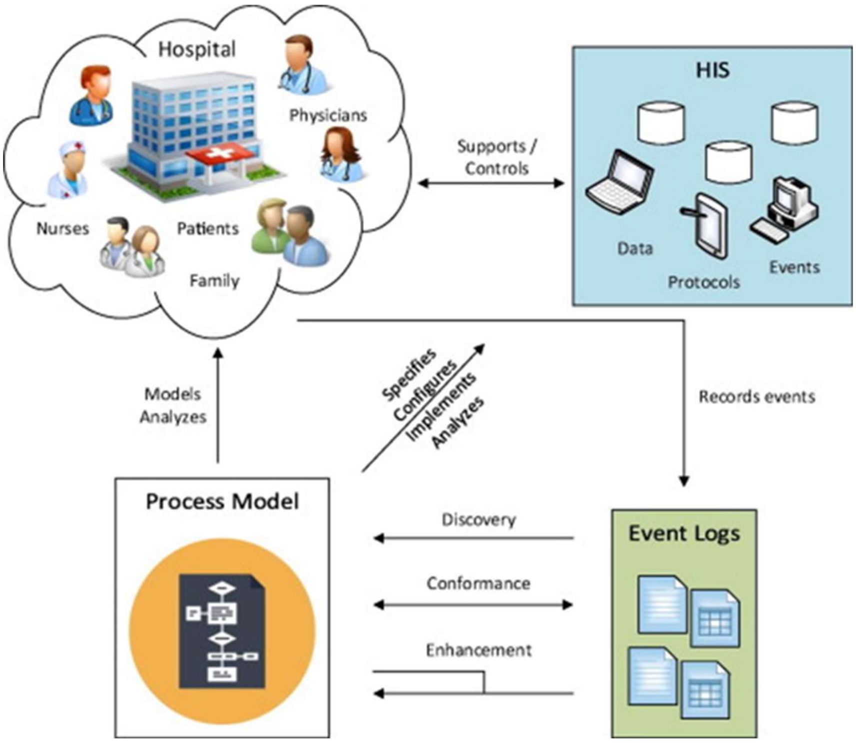
3) The SAIL datasets

The SAIL database was accessed via a dedicated gateway to secure research platform. Data was provided for all patients with an MSK event. Three primary datasets were used, these consisted of the Welsh Longitudinal General Practice (WLGP) dataset, the Patient Episode Database for Wales (PEDW) and the Welsh Demographics Service (WDS). Each GP practice uses a clinical information system to maintain an EHR at patient level. The WLGP dataset contains data on patient symptoms, test results, diagnosis, prescribed treatment, medications and referrals. The PEDW dataset holds data on a patient’s stay in hospital. This information relates to an individual hospital admission and includes data such as admission details, duration, diagnoses and operations. The WDS dataset was used to ensure that all patients were permanently resident in Wales at the time of their care. It includes information on the patient’s address, age and GP. Ethical approval for the study was granted on 01/10/2018 after submission of the SAIL Information Governance Review Panel application form.

1.5 Research landscape

According to Coiera [36], health informatics is the study of healthcare information systems and how they support processes in healthcare. One aspect of health informatics is the use of data from health information systems to generate insights that can improve healthcare provision. Process mining can help generate these insights. The diagram in Figure 1.3, reproduced from Rojas et al. [37], is used to describe the research landscape for this work. This research uses data from healthcare information systems to create event logs in order to perform process and data science. Methodological insights gained from this science may be used by process and data scientists working within the healthcare domain to help benefit patients, particularly those suffering with musculoskeletal conditions.

Figure 1.3 Landscape for this research



HIS represents any healthcare information system where healthcare data is captured and stored. Event logs and process models are produced in order to study the healthcare domain.

Health informatics as a science has a multitude of additional insights that may be relevant such as the psychosocial elements and human computer interaction. It is a broad and rich discipline that can provide many contributions. The contributions made by this research programme to the field of health informatics is through process-aware information systems.

1.6 Thesis structure

A brief description of the remaining chapters is provided below. The central structure of the thesis is based upon the three studies in chapters 5, 6 and 7, which reflect the order of experimentation performed during this research programme.

This chapter provides the reader with an overview of the main terms used within this thesis and introduces the problem domain by describing some of the current issues. An illustrative example is provided using Jack, a 50 year old male office manager. The hypothesis, along with primary research questions and objectives are stated before introducing the three data sources and their associated datasets that are to be used during the three studies. Finally, the research landscape is described.

Chapter 2 Healthcare background

This is the first of the two background chapters. Chapter 2 discusses the current healthcare literature in order to provide context and an understanding as a basis for this study. First, healthcare and clinical coding systems are discussed, before looking at the three different ways that are used in this research to provide a better understanding of healthcare. Finally, common MSK conditions used during the three studies are discussed.

Chapter 3 Technical background

Chapter 3 completes the background by reviewing and discussing various technologies appropriate to this research programme. Different aspects relating to relational and non-relational types of database management systems are presented, before exploring the field of data science. The three modelling notations used within this thesis are explained before discussing process mining. Literature relating to the three types of process mining: process discovery; conformance checking; and enhancement are discussed. Process mining methodologies and software tools are explored. The main literature reviews on process mining in healthcare are discussed, before looking in more detail at some of the challenges. The search criteria and results from a literature search into process mining in MSK is presented before finally summarising both the healthcare and technical backgrounds.

Chapter 4 Methodology

This chapter explains the method used for the three studies and describes in more detail the three datasets. The methodology is an adaptation of the PM^2 project management method and has five main stages which are: Planning; Extract, Transform and Load (ETL); Mining and analysis; Evaluation; and Process improvement. Common aspects of how these stages have been implemented is explained before describing in more detail the MIMIC-III, MyPathway and SAIL datasets.

Chapter 5 Process mining MIMIC-III data for disease trajectories

This chapter presents the first of the three studies. As stated above, chapters 5, 6 and 7 are structured using the five stages of the methodology. Only the areas that differ from Chapter 4 are presented. This study applies process mining techniques to the MIMIC-III data, in order to attempt to reproduce a disease trajectory model published in Nature Communications by Jensen et al. [38]. Refinements to Jensen's method are proposed and described to illustrate how a process mining approach can help to advance the study of disease trajectories from data stored in EHRs. The chapter begins by introducing the problem domain along with the disease trajectory model developed by Jensen. Study-specific research questions are defined. The chapter progresses through ETL to Mining and analysis, where two disease trajectory models are presented and finally evaluated in Stage 4. As the work in this chapter purely reports on research, there is no process improvement stage. Impact and future work are discussed before the chapter is summarised.

Chapter 6 Process mining MyPathway data to identify patterns of care

This is the second of the three studies. Here, process mining techniques are applied to data collected from the MyPathway mobile application in order to analyse different patterns of care, determined by patient reported health outcomes. Areas are identified where changes could be made to the MyPathway system in order to improve data collection for future process mining studies. The chapter begins by introducing the problem domain. The professional relationship with the data provider is explained and study-specific research questions are identified during the planning section. The ETL section is particularly large, as it reflects the many challenges encountered. Stage 2 is divided into the following sub-sections: an overview of the MyPathway mobile application; the management of MSK patient pathways at STHT; the definition of health outcomes using PROMs; data extraction; and data transformation and loading. During the mining and analysis stage, the data is first characterised before applying process discovery and analysis techniques. The method and results are evaluated before process improvements and future work are discussed.

Chapter 7 Process mining SAIL data for knee pain surgery pathways

Chapter 7 presents the third and final study. This chapter presents how an expert-defined reference model for a knee pain surgery was developed. It describes how this model was used to test the compliance of the real-life data SAIL and to help identify data cleansing issues, before using this data to generate episode statistics. Process mining techniques were used in order to: 1) discover the preliminary rules associated with a knee pain surgery pathway; 2) check compliance and identify data cleansing issues; and 3) produce statistics on time intervals between types of surgery and frequencies of different pathways travelled. The chapter begins by introducing the problem domain. An illustrative example of the knee pain patient pathway is presented by continuing the journey of Jack. Within the planning stage study-specific research questions are identified. Stage 2 describes the steps involved in ETL. The process mining and analysis steps are described before some episode statistics are presented. The work is evaluated and future work is discussed.

Chapter 8 Discussion and conclusions

Chapter 8 summarises the work carried out in this thesis and reflects on the method used for the three studies. Each of the studies are discussed in order, by considering their positive aspects, limitations and conclusions. Overall conclusions from this research programme are stated along with a discussion on how and whether the primary research questions have been answered. Contributions to knowledge, impact from this work and considerations for healthcare system development and process mining research are specified before providing suggestions for future work and a final summary.

Chapter 2
Healthcare background

Chapter 1 provided an overview of this thesis and introduced some of the challenges faced when needing to analyse routinely collected healthcare data. This chapter builds on the introduction by reviewing and discussing the relevant healthcare literature in order to provide a contextual background on a range of healthcare problems and challenges addressed in the thesis.

The problem domain decomposes into a number of healthcare areas. These areas include: healthcare systems; clinical coding standards; how change can be measured within healthcare; and common MSK conditions, relevant to this thesis.

2.1 Healthcare systems

Healthcare is the prevention, maintenance or improvement of health via the diagnosis and treatment of disease, injury, illness, and other physical or mental conditions. Healthcare in the UK is managed through primary, secondary and tertiary care services. Primary care is usually the first point of contact for people requiring healthcare and may be delivered by a GP, dentist, optician or pharmacist [39]. A person is often referred to secondary care, sometimes called hospital or community care, by a primary care practitioner (PCP). Secondary care is often received in short-term episodes, where more specialist attention is required for example a consultation, a course of physiotherapy or an elected surgical procedure [40], [41]. Referral to tertiary care is usually via a PCP where patients may receive highly specialised treatment for often rare or complex disorders such as transplants, neurosurgery and mental health services. Healthcare services are delivered by healthcare professionals which may include doctors, nurses, therapists, pharmacists, dentists and physical trainers.

Healthcare systems are developed in the historical and political context of each country. Most national health systems evolve to meet the challenges of demographic, economic, and epidemiological change along with changing technology in health [42]. Ideally each country would have universal access to comprehensive prepaid medical care, though unfortunately this is not the case. According to the World Health Organisation (WHO) every healthcare system has three main categories of stakeholder: government organisations and agencies; health service providers; and patients or service users [43]. All these stakeholders will influence its design. How a healthcare system is funded can influence the way it is used and subsequently impact on the information stored. For example, GP practices throughout the UK are financially rewarded through the Quality and Outcomes Framework (QOF) [44] for certain coded diagnoses within EHR systems. Many healthcare systems around the world have adopted the EHR. It can be described as the most common type of computerised healthcare information system and is a digital history, using a standardised format, of patient's health and healthcare that can be shared by different healthcare professionals [45]. Typical types of information within an EHR would include symptoms, diagnoses, medical and family history, examination and test results, procedures and treatments and medications. Globally, evidence exists that EHR systems have many advantages including quality of patient care, increased efficiency, an increase of patient trust and cost savings [46]–[48]. However, many are wary that they may fail or under deliver on the benefits initially promised [46], [49], [50]. The NHS is the healthcare system used in the UK and consists of NHS England, NHS Scotland, NHS Wales and Health and Social Care in Northern Ireland. Some of the most popular EHR systems in the UK include EMIS Health [51], SystmOne [52] and EPIC [53]. In the UK, almost total coverage for lifelong EHRs in primary care [54] has achieved. Though despite several attempts and due to many challenges, progress in secondary care has been much slower [55]–[57].

EHR systems are incredibly complex due to extreme diversity in their data and processes [58]. A process may be defined as a group of related activities that serves a common goal [59]. Healthcare processes can be described as a set of activities included in the diagnosis, treatment and prevention of a disease, with the intention of improving a person's health [60]. There are several guidelines used in healthcare to help standardise healthcare processes. The main source used in England is the

National Institute for Health and Care Excellence (NICE) [61]. NICE is part of the Department of Health, serving NHS England and NHS Wales. It provides guidance to help standardise health and social care for both clinical and economical purposes.

2.2 Clinical coding standards

The data stored within EHR systems is often recorded using clinical codes. There are a range of clinical coding standards used worldwide. This section will discuss coding standards relevant to this thesis.

The International Classification of Diseases (ICD) is the global standard diagnostic classification that is used for clinical and research purposes [62] and maintained by the WHO. The system maps a wide variety of signs, symptoms, abnormal findings, complaints, social circumstances and external causes of injury or disease to generic categories of similar diseases, assigning an alphanumeric code three to six digits long. The diseases are arranged into chapters. An example ICD-10 code is 'M17' which has the term 'Osteoarthritis of knee'. The letter 'M' relates to Chapter 13 which covers diseases of the musculoskeletal system and connective tissue, with specific disease codes ranging between M00-M99. The first version was introduced in 1893, ICD-9 in 1978 and the current version ICD-10, was introduced in 1992. Version 10 extended Version 9 by including significantly more diagnosis and procedure codes. It is possible to map between Version 9 and 10 codes. In 2018 ICD-11 was released and this version is planned to become operational in January 2022.

The Classification of Interventions and Procedures version 4 (OPCS-4) [63] is the system used to classify procedures within the NHS. The system assigns a four-digit alphanumeric code for operations, procedures and interventions performed during day case surgery, in-patient stays and some out-patient treatments within NHS hospitals. The first digit is an alpha character which relates to the chapter of the procedure, for example is 'W40.1'. Chapter 'W' contains procedures relating to 'Other bones and joints', the term for this code is 'Primary total prosthetic replacement of knee joint using cement'. The system was first published as the Classification of Surgical Operations in 1987 by the Office of Population Censuses and Surveys (OPCS). The

4th revision was released as OPCS-4.2 in 1992, before responsibility was passed to the NHS Information Authority in 1999. From then, there have been numerous revisions until the current version, OPCS-4.9, in April 2020.

Read codes are a clinical terminology system used within general practice in the UK [64]. Read codes were developed by the GP, Dr James Read in the early 1980's and have been used by the NHS in primary care since 1985. The system supports a wide variety of clinical coding for clinicians to record patient findings and procedures in electronic health and social care systems. Between 1988 and 2018 Read codes have evolved through three major design changes. The first version contained four-digit alphanumeric hierarchical codes. In 1991 Version 2, CTV2, was released due to a requirement from the NHS to provide a direct cross-mapping to both the ICD-9-CM and OPCS-4 systems. These codes retained the same technical properties though the length was extended to five digits. Later a drug and appliance dictionary was added. The final release was in April 2018, as by April 2020 the entire English health system was required to migrate to the SNOMED CT [65] system.

Within this thesis the following coding standards are used: Chapter 5, MIMIC-III data is recorded using the ICD-9 format and data used in the Jensen cardiovascular disease (CVD) trajectory model is recorded using the ICD-10 format; Chapter 7, SAIL data is recorded using the ICD-10, CTV-2 and OPCS-4 formats.

2.3 Understanding health

The overall aim of most health care is to improve health. More specifically, this may aim to: improve population health; reduce the cost of healthcare; improve efficiencies within the healthcare system; and improve the healthcare experience for the individual. In order to make healthcare improvements, the healthcare data must first be understood. There are various methods that can be used to assist in the understanding of health. The three methods that have been used in this thesis are:

- Creating disease trajectory models

- Comparing Patient Reported Outcome Measures (PROMs) at different points in time
- Comparing features within care pathways

The three sub-sections below explore and describe the current related literature.

2.3.1 Disease trajectories

There are many different applications of the term ‘disease trajectory’ [66]–[68]. Jensen et al. in their study using Danish population-wide registry data [38] described a disease trajectory as an ordered series of diagnoses, whereas Murray et al. expressed a disease trajectory as the progressiveness of physical health deterioration over time [69]. Murray separated disease trajectories into three kinds: short period, where the health declined in less than 24 months; long-term limitations, where the decline took between two to five years; and prolonged dwindling, where the decline took between six to eight years.

Creating disease trajectory models can help to inform epidemiologists and disease biologists [70]–[72] when measuring change. When used with EHR data, they are a powerful way to understand the complexities of relationships between diseases [73], [74]. Disease trajectory models are frequently represented as directed and often acyclic graphs. The nodes represent diseases and the directed arcs represent the progression between diseases. The removal of cycles renders a progressive trajectory where repeating diseases are not permitted. This is often a simplification of reality where a clinical decision must be made. It is sensible to base the decision on whether the disease group mainly contains chronic or acute diseases, as seen with Murray. For example, an acyclic graph should be created for a group of CVDs due to their long-term nature, as any recording of a repeated diagnosis is likely to be related to an earlier instance, rather than a new occurrence of the disease. The decision of whether to include repeating diagnoses must be made when extracting the EHR data. If the decision is made to include repeating diagnoses for a patient, it must then be decided whether to include those repeating diagnoses on the model. Jensen, included repeating diagnoses when extracting the data, as they did not want to make the assumption that the first time a diagnosis appeared in the data, was the first time it appeared in the patient [75].

Creating disease trajectory models from EHR data is challenging and time consuming. Various methods have been used including data mining [76]–[79], data-driven [77]–[79] and network-based approaches [83], [84]. There is no recommended standard evident within the literature for creating disease trajectory models. In Chapter 5 disease trajectory models are created using a process mining approach.

2.3.2 Patient Reported Outcome Measures

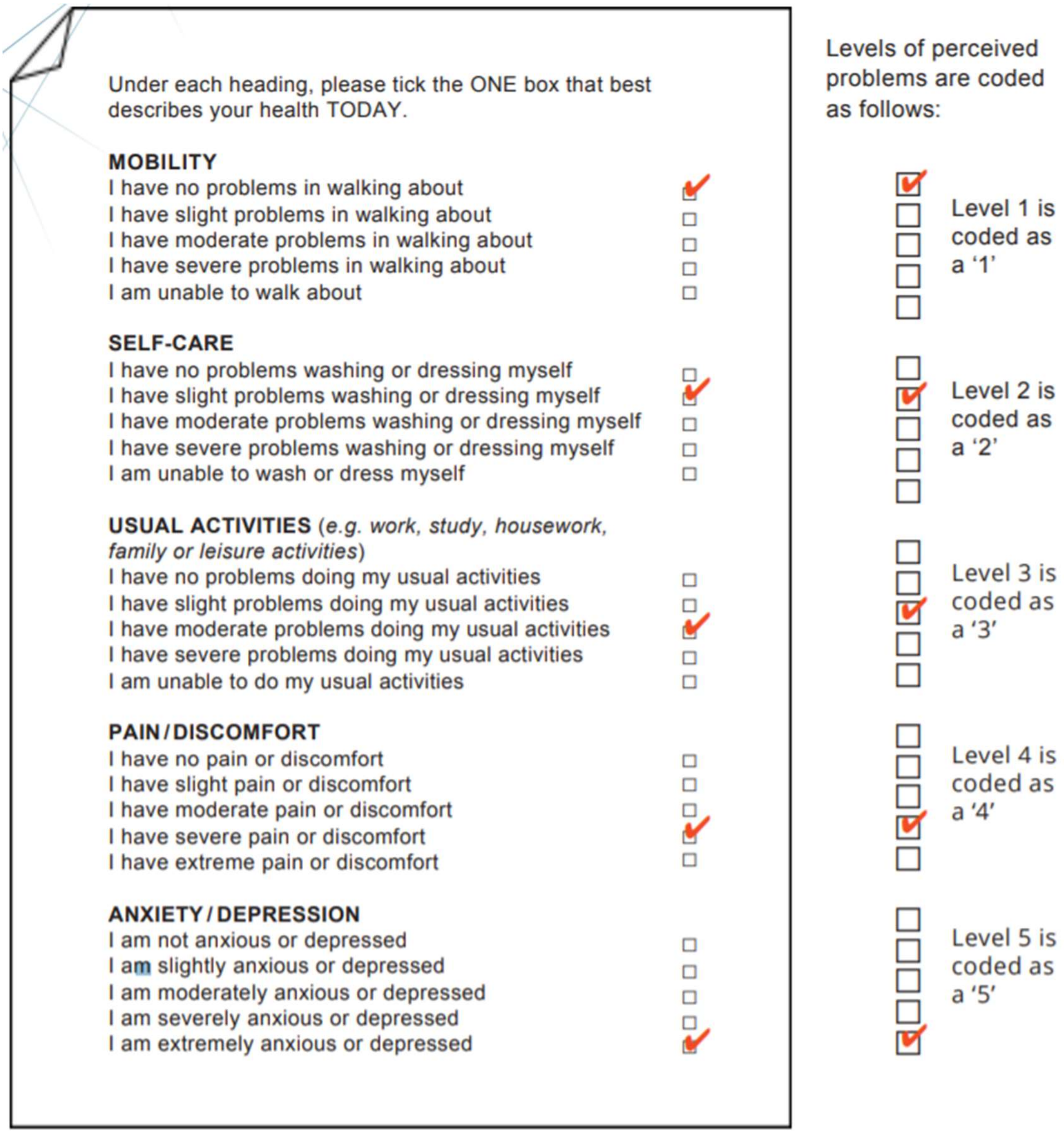
Patient Reported Outcome Measures are tools or instruments, often in the form of questionnaires, used to record in a structured way patient reported outcomes. PROMs can capture condition-specific information, an example is the Oswestry Disability Index [85] or more general quality of life information, such as the SF-12 [86]. They are used in many situations for example population health surveys, clinical trials, observational studies and routine data collection for healthcare services [87]. Increasingly PROMs are being used to assess the effectiveness of elective orthopaedic procedures [88], [89]. Tew et al. in a recent study [86] on health-related quality of life (QoL) trajectories following total knee replacement (TKR) surgery concluded that there is scope in translating routinely collected PROMs in order to improve shared decision making. However, they stated that more research is needed to identify appropriate approaches of its implementation to guide clinical care and maximise patient health outcomes.

Due to technological advances, the distribution, collection and analysis of PROMs data is now often inexpensive and feasible [90]. When collected over time they are useful to help understand, from a patient’s perspective, the effect of healthcare interventions. PROMs enable evaluations and comparisons to be made across healthcare providers and services [87]. By helping clinicians to provide more patient-centred care [91], they are believed to have a positive effect on a patient’s experiences and outcomes [92], [93]. The NHS began collecting PROMs for hip and knee replacement surgery in 2009 with the national PROMs programme [94]. Systematic reviews on PROMs within MSK include Fennelly et al. where PROMs in advanced physiotherapy practice [95] were reviewed and Ramkumar et al. who reviewed PROMs after total knee arthroplasty [96]. These reviews concluded that PROMs are widely used across a range of MSK disorders, though gaps do exist. Reliability, validity, responsiveness [97] and interpretation guidelines such as Minimal Clinically Important Difference (MCID) must be considered [98], [99]. The PROM instrument used for this research is the generic EQ-5D-5L which is widely used in healthcare studies [100], [101] and also within MSK research [102], [103].

2.3.2.1 The EQ-5D

The EQ-5D [104] is a simple, generic measure of health-related quality of life (QoL) for both clinical and economic evaluation, created by the EuroQol Group [105] in 1988. It has evolved over the years with the latest version being the EQ-5D-5L [106]. There are two sections, the first on page one using the descriptive system and the second on page two using a visual analogue scale (VAS). An example of a completed section one is presented in Figure 2.1.

Figure 2.1 Example of a completed EQ-5D using the descriptive system [106]



Under each heading, please tick the ONE box that best describes your health TODAY.

MOBILITY
I have no problems in walking about ☑
I have slight problems in walking about □
I have moderate problems in walking about □
I have severe problems in walking about □
I am unable to walk about □

SELF-CARE
I have no problems washing or dressing myself □
I have slight problems washing or dressing myself ☑
I have moderate problems washing or dressing myself □
I have severe problems washing or dressing myself □
I am unable to wash or dress myself □

USUAL ACTIVITIES (*e.g. work, study, housework, family or leisure activities*)
I have no problems doing my usual activities □
I have slight problems doing my usual activities □
I have moderate problems doing my usual activities ☑
I have severe problems doing my usual activities □
I am unable to do my usual activities □

PAIN / DISCOMFORT
I have no pain or discomfort □
I have slight pain or discomfort □
I have moderate pain or discomfort □
I have severe pain or discomfort ☑
I have extreme pain or discomfort □

ANXIETY / DEPRESSION
I am not anxious or depressed □
I am slightly anxious or depressed □
I am moderately anxious or depressed □
I am severely anxious or depressed □
I am extremely anxious or depressed ☑

Levels of perceived problems are coded as follows:

☑ □ □ □ □ Level 1 is coded as a '1'

□ ☑ □ □ □ Level 2 is coded as a '2'

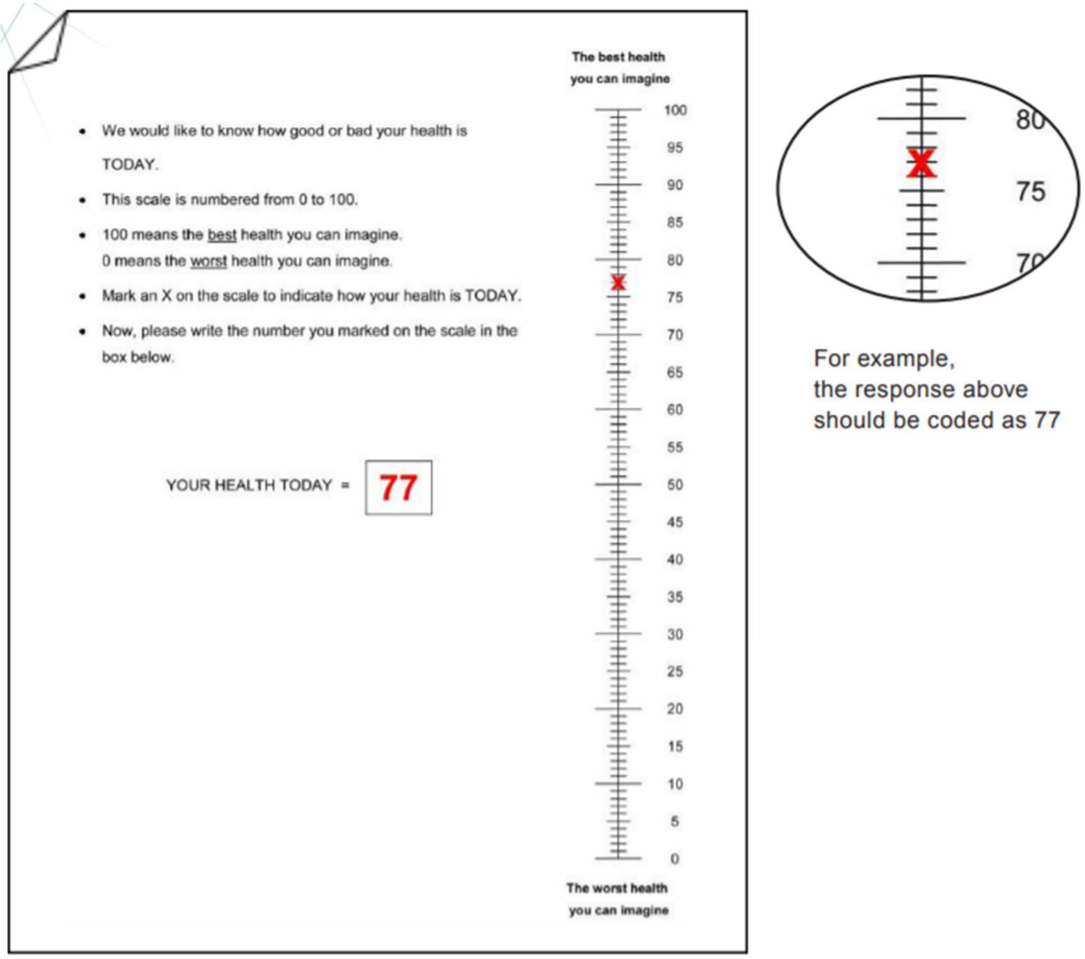
□ □ ☑ □ □ Level 3 is coded as a '3'

□ □ □ ☑ □ Level 4 is coded as a '4'

□ □ □ □ ☑ Level 5 is coded as a '5'

The descriptive system has five dimensions consisting of mobility, self-care, usual activities, pain/discomfort and anxiety/depression. Each dimension has five associated severity levels where the patient rates how they feel that day ranging from 'no problems', coded as '1' to 'extreme problems', coded as '5'. The five values together form the health state. The health state for the example above is '12345'. An example of a completed section two is presented in Figure 2.2.

Figure 2.2 Example of a completed EQ-5D using the VAS [106]



- We would like to know how good or bad your health is TODAY.
- This scale is numbered from 0 to 100.
- 100 means the best health you can imagine.
  0 means the worst health you can imagine.
- Mark an X on the scale to indicate how your health is TODAY.
- Now, please write the number you marked on the scale in the box below.

YOUR HEALTH TODAY = 77

The best health
you can imagine
100
95
90
85
80
75
70
65
60
55
50
45
40
35
30
25
20
15
10
5
0
The worst health
you can imagine
X
80
75
70

For example,
the response above
should be coded as 77

In Section 2 the patient places a cross on a scale of zero to 100, with zero being the worst imaginable health state and 100 being the best. When analysing the results for the descriptive system it is important that no arithmetic value is associated with these numbers. Given there are five dimensions with five levels of severity within the descriptive system, there are 5^5 or 3,125 possible health states. Numerous studies have been undertaken to assess the reliability and sensitivity [107]–[109] of the EQ-5D.

Information may be reported in one of these three ways [100] as: a health profile; a single health index; or a VAS. The health state from the descriptive section may be left unaltered and presented as a health profile, retaining all data for each of the domains. An example of a person's health profile could be 33112, this would equate to them being confined to bed, unable to wash or dress themselves, no problems with usual activities, no pain and moderately anxious. EuroQol state that reporting the information using health profiles is the preferred method unless there is a good reason not to [110]. Alternatively, a single health index uses value sets [111] to determine a summary score. Value sets are a set of weights applied to each of the levels and are created for different countries by selecting residents to place a value on the different health states. However, reporting using a single health index can cause bias [112] due to the introduction of social values. In addition, valuable information is lost when the data is aggregated [113], [114]. This method would usually be used in health economics analysis where a single number between 0 and 1 is required. Finally, the VAS data requires no pre-processing and can be represented as a quantitative measure of the patient's valuation of their own health status. Although this data can be informative, capturing broader health factors [115], it is often disregarded during analysis. It is interesting how patients often rate themselves as having no problems in the domains, though do not enter a value of '100' for the VAS [116].

It would be helpful to know if the severity of the scores at baseline influenced a patients' outcome. Kolotkin et al. [117] reported that health related quality of life improvements are greater for individuals who have more severe scores at baseline. Whereas Stratford et al. [118] deems this unlikely, as they found patients with more severe pain at baseline required to see more of an improvement for the change to be regarded clinically important. Responses from the descriptive section, in the form of health profiles are used in Chapter 5 to determine a patient's health outcomes.

2.3.2.2 The Paretian Classification of Health Change

Although the preferred method for reporting of EQ-5D data is by health profile, as it is constructed from five ordinal variables, it can be difficult to determine the overall change for a patient. When performing non-financial analysis, one of the simplest methods, without creating a single summary score, is to use the Paretian Classification of Health Change (PCHC) [114]. In summary, this method compares a matched pair of questionnaires to determine a patients' health outcome by using the following rules:

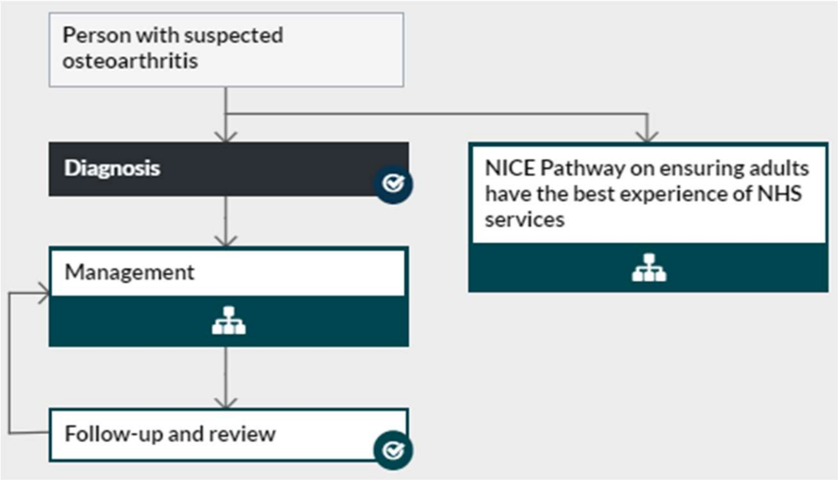
- A health improvement is where a profile is better in at least one dimension, and no worse in any other.
- No change is where both health profiles are the same.
- A mixed health profile is where there has been an improvement in at least one dimension and a decline in at least one dimension.
- A health decline is where a profile is worse in at least one dimension, and no better in any other.

Although the PCHC provides a simple reporting method, it is of limited use when the 'mixed' category in the dataset is dominant [114]. A recent study that evaluated the usefulness of the EQ-5D using the PCHC among patients with long-term conditions [119] found that the method could identify subgroups that showed overall health change and helped to identify needs for more tailored interventions. The PCHC is the method chosen to identify health change in Chapter 6.

2.3.3 Care pathways

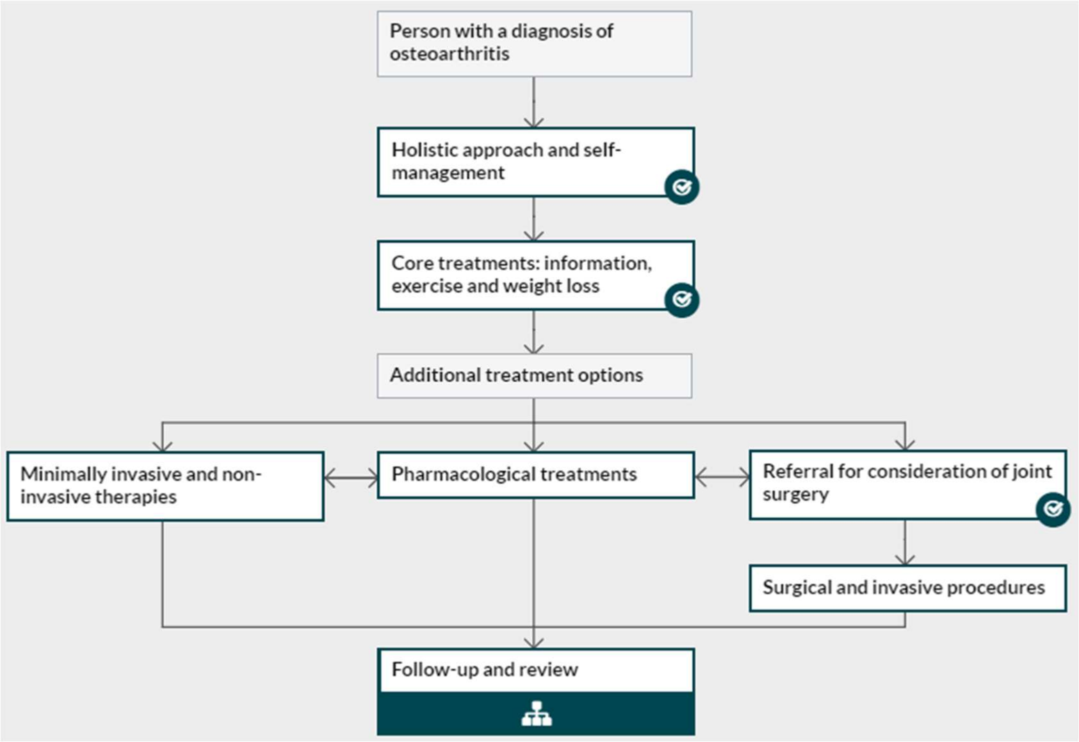
With an increasing need for more efficient healthcare, there is an emphasis on establishing evidence-based [120] best practice guidelines. These guidelines, which are often implemented by interdisciplinary teams, can help to establish a strategy for the prioritisation of care and reduction of variation [121]. Care pathways, also known as critical or clinical pathways [121] are a way of viewing the implementation of these best practice guidelines [122], with the aim of defining an optimal treatment process which can be used to improve the quality of patient care and reduce healthcare costs [121]. Clinical pathways are often hospital and procedure specific, developed to create the best process for patient-centred care within a specific institution [123]. However, as mentioned in Section 2.1, NICE produce care pathways for general use across England and Wales. These care pathways relate to specific conditions, Figure 2.3 presents the high-level care pathway for osteoarthritis.

Figure 2.3 NICE high-level pathway for osteoarthritis [124]


Person with suspected osteoarthritis
Diagnosis
NICE Pathway on ensuring adults have the best experience of NHS services
Management
Follow-up and review

These pathways are interactive and activities containing the flowchart symbol can be further expanded as in Figure 2.4.

Figure 2.4 NICE pathway for the management of osteoarthritis [125]


Person with a diagnosis of osteoarthritis
Holistic approach and self-management
Core treatments: information, exercise and weight loss
Additional treatment options
Minimally invasive and non-invasive therapies
Pharmacological treatments
Referral for consideration of joint surgery
Surgical and invasive procedures
Follow-up and review

When activities are selected detailed textural information is displayed.

In 2019 a literature review of care pathway modelling was carried out [126] by Aspland et al. Aspland states that care pathways have been increasingly developed and used since 1985. Different methods used to create care pathways include BPMN, simulation, data mining and more recently process mining. In this review the studies were evenly distributed amongst acute and chronic conditions, though studies including care pathways for surgical procedures only totalled 11 percent [126].

Button et al. found that though a number of related evidence-based guidelines exist, MSK care pathways are variable and inconsistent, with little consideration given to the organisation of care [127]. This could be partly due to the number of options open to GPs when conservative treatment is unsuccessful, such as referral to another GP with an interest in MSK conditions, an advanced nurse practitioner, a physiotherapist or an orthopaedic surgeon [128]–[130]. Though care pathways are a useful tool, capturing and measuring the differences between patient traces left in EHRs when patients have veered from the norm can be used to provide valuable insights [130]. Huang comments that bringing order into the chaos of care pathways is not recommended as many existing techniques for pathways analysis focus on aggregated traces and therefore lose the high degree of variability within patient care. Carol et al. agrees and states that patient diagnoses and treatments often deviate considerably from the standard clinical pathways [131] and that analysing these deviations may result in improved quality of care.

2.4 Common musculoskeletal conditions

Musculoskeletal conditions affect the locomotor system which includes the bones, joints and associated soft tissue such as tendons, ligaments, muscles and nerves [4]. They may be short-lived due to an injury or be a long-term condition which could last for life. One in five people [4] suffer from MSK related pain during their life-time. The pain is typically persistent and often restricts mobility. In advanced cases, a person's ability to work and socialise may be reduced, often resulting in depression and an increased risk of developing further long-term conditions [132]. Arthritis is a term often associated with MSK conditions and means inflammation of the joints, though is commonly used by doctors to describe any condition affecting the joints.

There are over 200 different MSK conditions [6] with the most common being osteoarthritis (OA). The WHOs Global Burden of Disease Study (GBD) [133] is the most comprehensive worldwide epidemiological study to date, examining epidemiological trends of 291 diseases across 187 countries. The study discusses morbidity and mortality and identifies risk factors worldwide [5], [134]–[136]. The GBD study showed MSK conditions to be the greatest cause of disability in the UK and second worldwide with the main contributors being lower back and neck pain and OA. Between 2015 and 2016 in England, MSK conditions accounted for approximately 15 percent of the total admitted patient care [137], the highest of all 21 healthcare chapters and over 25 percent of surgical interventions in the NHS [132]. The following four sections provide an overview of the four MSK conditions used in this research programme.

2.4.1 Osteoarthritis

Osteoarthritis can be a debilitating condition, often referred to as *wear and tear*. It typically presents in the spine, hips, knees, hands and feet, affecting approximately 13 percent of the UK population [138]. The NICE guidelines state that a diagnosis of primary OA should be given if alternative conditions have been excluded, the patient has activity-related knee pain, is aged 45 years or over and suggestive clinical features exist. Clinical features include a history of activity-related joint pain with either no morning stiffness or morning stiffness lasting no more than 30 minutes [139].

OA is caused by a degradation of joint cartilage and results in the rubbing together of bones often causing inflammation and over time structural changes to the joint. This structural change is due to a response to chondrocytes in the articular cartilage and the inflammatory cells in the surrounding tissues. The release of enzymes from these cells destroys the articular cartilage, exposing underlying subchondral bone which often results in sclerosis, particularly in load-bearing joints such as the knees and hips. The joint space is progressively lost over time and bony growths known as *spurs* or *osteophytes* may develop causing damage to the soft tissues around the joint, resulting in increased pain and reduced movement.

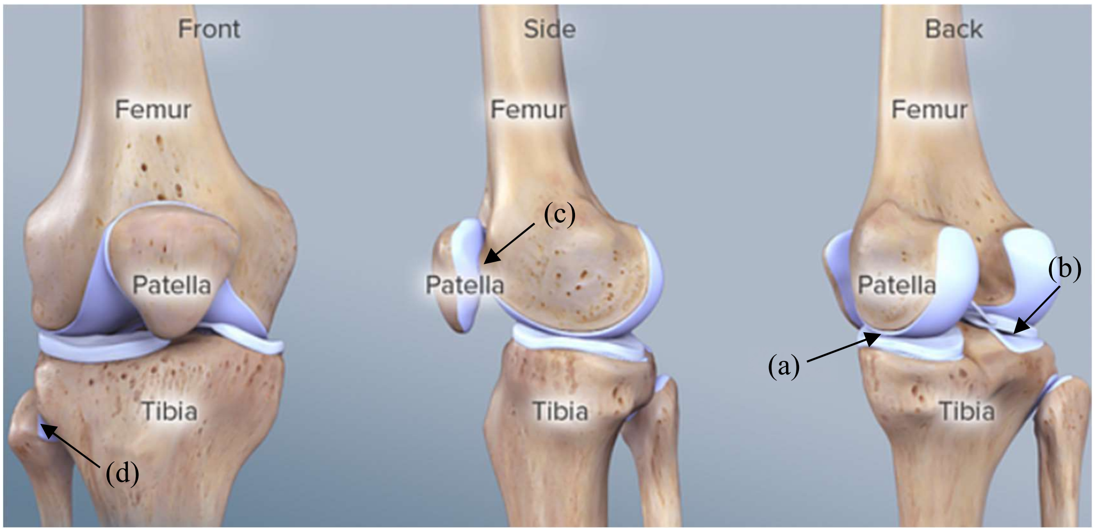
OA has a variety of causes and outcomes often differing by joint, though usually with a common pathology in the specific condition [140]. Recent studies however have shown considerable heterogeneity between groups of patients in relation to the progression of the disease [141]–[144]. The main risk factor is age [145] with the highest risk group being between 45 and 75 years [139], [146]. Generally, women are more likely than men to develop the condition, this is often attributed to their smaller muscles, increasing effort on the joints [147]. Obesity puts additional strain on weight bearing joints, especially the knees and hips [3]. High bone density increases the risk for OA of the hip, hand and knee, whereas low bone density is linked to fast progression in the knee and hip [138]. During childhood the size and shape of developing bones determine the biomechanics of the joint and can influence the risk of developing OA in adulthood [148]. Diagnosis of OA is usually made due to the presentation of symptoms followed with a physical examination by a GP. If a diagnosis is unclear, blood tests and imaging can be performed to exclude other conditions, though these methods can be imprecise which may result in an inconclusive diagnosis [149]. Core treatments include education for effective self-management, activity and muscle strengthening exercises and weight loss if required. In addition to these core treatments, medication may be prescribed for pain relief. In the first instance paracetamol and topical Non-Steroidal Anti-Inflammatory Drugs (NSAIDs) are recommended. Where these prove insufficient, opioid analgesics may be added. For moderate to severe OA, intra-articular corticosteroid injections can be administered along with existing medications. In severe cases when other treatments are ineffective, often in late stage OA of the knee or hip, joint replacement surgery may be an option [139].

OA is a subset of joint pain. For the purpose of this thesis the focus is on knee and back pain. There are also other causes as is described in more detail below.

2.4.2 Knee pain

The knee is an extremely complex joint [150] consisting of (a) the medial tibiofemoral, (b) the lateral tibiofemoral, (c) the patellofemoral and (d) the proximal tibiofibular joints which can be seen in Figure 2.5.

Figure 2.5 Bones of the knee joint [151]



Cartilage is represented in purple in Figure 2.5. Articular cartilage covers the bones where they meet and is a smooth, extremely strong and flexible fibrous tissue. It cushion the bones and allows them to glide over each other as the knee bends and straightens. Each knee has two menisci which are thick pads of cartilage situated between the femur and tibia. Knee meniscus acts as a shock absorber and helps to stabilise the knee. Four major ligaments (not shown in Figure 2.5, which are bands of strong tissue, connect the femur to the tibia. These ligaments provide stability in all directions. The knee’s largest tendon is the patellar tendon, which connects the thigh’s quadriceps muscles to the patella and down to the front of the tibia. Bursae surround the knee and are tiny, synovial fluid-filled sacs which reduce friction between the bone and soft tissue. A healthy knee requires the surrounding muscles to be strong and flexible, these muscles include the quadriceps and the hamstrings at the front and back of the thigh and the lower leg muscles.

Due to this complex anatomy knee pain is common especially in adults over 50 years of age [152]. The main cause is osteoarthritis, though other causes include injuries, tumours, referred pain and bursitis [153]. The knee is the most frequently affected joint with OA and represents the leading cause of disability in the adult population [154]. A number of evidence-based guidelines exist for OA of the knee [139], [155], [156] and as such knee pain is usually diagnosed by a GP upon examination. OA of the knee is typically bilateral and symmetrical. Unilateral OA of the knee is usually secondary to an existing trauma or disease [157]. In addition to the diagnosis criteria for general OA, people suffering from OA of the knee may experience difficulty walking, climbing stairs and progressive aching when sitting that is relieved by standing. On examination there may be bony swelling and joint deformity, small to moderate amounts of fluid, warmth, muscle wasting and weakness, restricted and painful range of movement and crepitus which is a grating sound or sensation.

In addition to the risk factors for general OA, there is extensive evidence that obesity is a major risk factor [158]–[161]. Others include knee-straining work and sport [153]. The global prevalence of OA of the knee is higher in women [162], peaking at 50 years and rises with age [137]. The lifetime risk of developing symptomatic OA of the knee in the UK has been estimated at 45% [163]. Lifetime risk is a statistic used to describe person-level risk, which is the probability of developing a condition over the course of a lifetime. In late-stage OA of the knee, when all other treatments have become ineffective, total knee replacement (TKR) surgery also known as total knee arthroplasty (TKA) is an option.

2.4.2.1 Knee pain surgery

Chapter 7 examines four types of knee pain surgery, these are: arthroscopy; primary total knee replacement; revision of total knee replacement; and attention to total knee replacement. A brief description of these surgery types is given in the sub-sections below.

2.4.2.1.1 Arthroscopy of the knee

Arthroscopic knee surgery is a type of keyhole surgery and is used to diagnose and treat knee joint problems. It is often used to assess the level of damage after an injury or from underlying conditions such as osteoarthritis when scans return inconclusive. It may also be used to treat damaged cartilage, to remove loose fragments of bone or cartilage or for joint aspiration when patients experience persistent joint pain [164]. Arthroscopic surgery has many advantages over open surgery due to the small instruments used, therefore requiring a smaller incision. These advantages include a lower risk of infection, no overnight stay, less post-operative pain and a faster recovery time.

During the procedure the patient will usually be under general anaesthetic, although sometimes a local anaesthetic or epidural is used. A small incision is made, measuring a few millimetres next to the knee joint where a thin tube with a camera (arthroscope) is inserted. It is sometimes helpful to fill joint with fluid to make it easier for the surgeon to see. If treatment is necessary additional incisions are made for the insertion of the tiny surgical instruments.

After the procedure the incisions are closed with surgical tape or stitches and dressed. The procedure usually takes between 30 and 45 minutes. Once the patient has recovered from the anaesthetic a physiotherapist will discuss post-operative exercises and provide the patient with crutches to use during recovery. The patient is usually discharged the same day, with a follow-up appointment in two weeks. Normal activities can usually be resumed within six weeks.

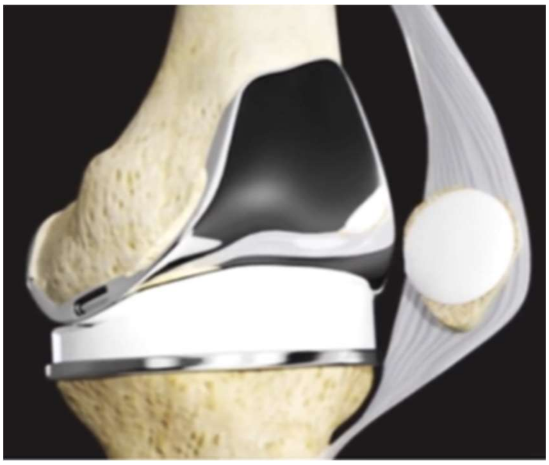
Arthroscopy of the knee is one of the most common surgical procedures worldwide with the number considerably increasing in the last 30 years [165]. However, opinion is divided, with most authors claiming the procedure to be inefficacious, especially when performed to treat osteoarthritis [166]–[170] and others disputing this claim [171].

2.4.2.1.2 Primary total knee replacement

Total knee replacement surgery has been routinely done for over 40 years is one of the most frequently performed surgical procedures worldwide [152], [172] with in excess of 100,000 per year in the UK [173]. The most common indication for the procedure is painful osteoarthritis of the knee (n=95%) with reduced mobility and quality of life [173], [174]. Though the decision to operate is influenced by many factors including patient and surgeon preference [173]. TKR surgery successfully reduces knee pain and improves long-term function in approximately 80% of cases [165], [166]. However, 20% of patients are disappointed with the result [172]. People are offered surgery when their knee pain interferes with their quality of life, though their state of health before the operation must be adequate to cope with both a major operation and rehabilitation. Patients of any age may be considered, although most surgeries are carried out on people between the ages of 61 and 76 [174]. It is important that the muscle surrounding the joint is as strong as possible before the surgery, and therefore if possible the patient should continue to exercise.

Total knee replacement surgery is performed when the articular cartilage is damaged over the entire surface of the knee joint, this includes the upper end of the tibia, the lower end of the femur and patella. The surgery takes between two and three hours and is performed under general or spinal anaesthetic. Using open surgery, the surgeon makes an incision down the centre of the knee before removing any damaged articular cartilage and bone. The upper end of the tibia is replaced with a flat metal plate and a plastic spacer. The lower end of the femur is replaced by a curved piece of metal and the patella is resurfaced with a plastic layer. These three parts are secured either with or without cement, an example can be seen in Figure 2.6.

Figure 2.6 Example of a total knee replacement [175]



An artificial joint is also known as a prosthesis. After surgery the patient will usually remain in hospital for between three and five days where a physiotherapist will provide them with crutches or a walking frame and knee strengthening exercises. It is important that the patient is mobile as soon as possible after the surgery. Normal activities can usually be resumed within six to eight weeks, though full recovery can take up to two years for scar tissue to repair and muscles to restore to full strength [176].

The use of patient reported outcome measures for the evaluation of TKR surgery has become more common. A recent systematic review showed three of the more popular condition-specific instruments to be the Western Ontario and McMaster Universities Osteoarthritis (WOMAC), the Knee Injury and Osteoarthritis Outcome Score, and the Oxford Knee Score (OKS), as well as the more general quality of life EQ-5D [177].

2.4.2.1.3 Revision and attention to total knee replacement

Despite the high success rate of primary TKR surgery, some patients require revision surgery, this is where the original prosthetic joint is replaced. The most common presenting sign is pain. A recent systematic review considering the longevity of knee replacements, stated that 82% of TKRs last for 25 years [178]. Other studies found that reasons for failure differed between early (within the first two years) and late revision (greater than two years). Overall, the most common reasons for revision surgery were aseptic loosening, periprosthetic joint infection, progressive arthritis and periprosthetic fracture [173], [179], [180]. Though, a recent study [181] looking at patient demographics found that early revision and re-revision surgery was carried out almost twice as often in younger patients under the age of 55.

Revision total knee replacement surgery is usually performed under general or spinal anaesthetic. Again using open surgery, the surgeon will make an incision down the centre of the patient's knee before removing the old prosthetic implant. If bone grafts are required, this is done before fitting the new knee components and closing the wound. The post-surgery process and timescales are similar to those for the primary TKR.

Attention to TKR surgery is required when a patient experiences problems, often stiffness, after their TKR surgery, though a revision is not be necessary at that point in time. In these cases the patient will attend an outpatient clinic for treatment. Stiffness after total knee replacement surgery ranges from 1.3 percent to 5.8 percent

[182] and is often caused by the formation of scar tissue. This stiffness can often be relieved by the manipulation of the joint whilst under anaesthesia.

2.4.3 Back pain

Back pain is the largest cause of disability in the UK [183] affecting around 10 million people and is the second most frequent cause of short–term absences after minor illnesses [184]. Back pain is a general term and often includes pain anywhere along the spine from the lumbar spine to the neck. Back pain may stem from an underlying condition such as an infection, tumour, osteoporosis, arthritis, fracture or structural deformity, or may be non-specific, meaning it has no known specific pathology. A study estimating the burden of MSK disorders in the community [185], reported that in the UK back pain is the most commonly reported site of MSK related pain for people under 65 years and affects approximately 33 percent of the adult population each year [183]. Often back pain is not serious and may be due to a soft tissue strain or inflammation, though it can be due to a chronic condition and reoccur over a life-time. The majority of patients with back pain are managed in primary care, where back pain accounts for 7 percent of all GP consultations [186]. Twenty percent of all MSK consultations are associated with the back, of which 14 percent are specific to the lower back [183]. Lower back pain (LBP) affects one in ten people [187] and is usually defined as pain or stiffness localised below the lower chest and above the gluteal folds [187]. LBP may or may not involve leg pain which is known as *sciatica*. Lower back pain is ranked highest in the world for causing disability, with neck pain fourth [133]. The term spondylosis is used for OA of the spine, usually the lumbar (lower) spine. With spondylosis the bones of the spine and the cushioning discs that separate the spinal bones become damaged. With spondylosis of the lumbar spine it is common to experience numbness or weakness to the legs and with spondylosis of the neck, pain or pins and needles in the shoulders or arms can occur. In severe cases spondylosis can result in the compression of the spinal cord causing damage, known as myelopathy.

The GBD study showed LBP is more common in men than women [188] and peaks at 80 years of age, whereas with neck pain women are at higher risk [5]. The major risk factors for back pain are obesity, depression, smoking and deprivation. People suffering from obesity are four times more at risk to develop back pain [189]. People suffering from depression are twice as likely to develop back pain [190] and smokers are approximately 50 percent more likely than non-smokers [191]. Finally, pre-retirement aged people between 45 and 64 years living in the most deprived areas are almost twice as likely to report back pain [192]. Management guidelines differ depending on the underlying cause and the area of the spine affected. As non-specific LBP does not have a known pathoanatomical cause, focus is to reduce the pain. Management involves education, analgesic medication, non-pharmacological therapies and timely reviews [193]. Unfortunately, the overuse of opioids and imaging is a widespread problem [193] and patients are often given injections with low evidence of their effectiveness [194]. Where traditional management is ineffective, surgical intervention may be an option, though referrals are increasing with a large amount of treatments having a poor evidence base [194].

2.4.4 Gout

Gout, also known as *gouty arthritis* is the most common type of inflammatory arthritis. Despite rising prevalence [195], poor management of the condition continues in many countries [196]. Recent reports show that prevalence varies considerably from between less than 1 to 6.8 percent [196], [197]. The highest reports are in Oceanic countries, especially within certain ethnic groups, such as Maori and Taiwanese Aboriginals [198]. Gout is characterized by increased levels of uric acid in the blood leading to a build-up of sharp, urate-containing crystals in the joints and soft tissues, often causing severe pain and disability [197]. Symptoms usually occur suddenly in the big toe, though other joints may include the ankle, knee, elbow, wrist and finger. The affected area often appears swollen, red, warm and tender. In 2015 the American College of Rheumatology (ACR) and the European League Against Rheumatism (EULAR) created an updated gout classification criteria [199].

The risk of gout increases with age and is more common in men than in women by a ratio of approximately 4:1 [5]. Other risk factors include: obesity; a high intake of alcohol; red meat; seafood and sugary drinks; plasma urate raising drugs; and a family history of gout. People with the following comorbidities are also at an increased risk: hypertension; hyperlipidaemia; diabetes; cardiovascular disease; chronic kidney disease; osteoarthritis; myeloproliferative disease; psoriasis; sickle cell anaemia; renal disease; and glycogen storage diseases. NICE guidelines state that for an acute attack, treatment with NSAIDs, colchicine or corticosteroids should commence within 24 hours from the onset of symptoms. At a four to six week follow-up they recommend discussing the use of a Urate-Lowering Therapy (ULT) such as allopurinol or febuxostat [200]. However, in 2017 a review of updated treatment guidelines stated that varying guidelines were followed with differences in the timing for starting and indications for receiving ULT [197]. Despite rising prevalence and available guidelines gout is often poorly managed with only one-third to a half of gout patients receive ULT [196], [198].

The work in this chapter will be summarised in Section 3.7.

Chapter 3
Technical background

Chapter 1 introduced some of the problems faced within healthcare for patients suffering from MSK conditions. Chapter 2 expanded on this by providing a background on the relevant healthcare literature and discussing further the range of healthcare problems and challenges. This chapter completes the background by exploring the technical literature and discussing the range of technologies that can be used to address these healthcare problems.

The solution toolset decomposes into a number of technical areas. These areas are: Types of database management systems; data science; process modelling; process mining, including the different types, process mining methods and software; process mining in healthcare. Process mining in the healthcare domain poses many different challenges, often due to the nature and complexity of healthcare processes. These challenges are introduced in Section 3.5.1 and explored further within the three studies in chapters 5, 6 and 7.

3.1 Types of database management system

Data within large information systems is stored in a database. A database management system (DBMS) is a software package designed to store, retrieve and manipulate data in a database. There are many types of DBMS, the main types are: hierarchical; NoSQL; object-oriented; and relational. Each type of DBMS has advantages and disadvantages and is chosen depending on the type of data it needs to store and functions it needs to perform. The two DBMS described below are used in the studies in chapters 5, 6 and 7.

3.1.1 Relational

Relational databases management systems (RDBMS) were first described by Codd in 1970 [201]. They are based on the relational model where data is grouped into relations or tables where data items, known as attributes, are ordered in rows and columns. Tables are connected using primary and foreign keys that join rows together using unique identifiers. The Structured Query Language (SQL) is the standard programming interface for creating, manipulating and querying a relational database. In 1994, Distributed Relational Database Architecture [202] was introduced which enables network connected relational databases to work together using SQL requests. Types of RDBMS include Microsoft SQL Server, Oracle database, MySQL and IBM DB2. The data sources used in chapters 5 and 7 are both relational.

Some of the advantages of this kind of database are that they: contain structured data that is easy to understand; have built-in data integrity; enforce constraints within the relationships; are familiar to most IT professionals; use the standard SQL; and have limitless indexing. However, there are also some disadvantages, which in many ways have been emphasised with the introduction of big data. Some of these disadvantages include: issues with concurrency and scaling; data adhering to the relational model (normalised) requires many joins which impacts on data retrieval speed; a rigid structure which cannot easily be altered when requirements change; and problems when working with semi-structured data. To circumvent some of these problems NoSQL databases were introduced.

3.1.2 NoSQL

The acronym NoSQL stands for ‘not only SQL’. It was first used in 1998 by Carlo Strozzi for his relational database that did not use SQL. While NoSQL databases have existed for many years, it was not until late 2000s, in the era of big data and high-volume internet and mobile applications, that they became more popular [203]. Around this time, the cost of storage considerably decreased. Cheaper storage led to more data being stored both on and off the cloud, often distributed across many servers, in structured, semi-structured, unstructured and polymorphic formats. In 2001, the Agile Manifesto [204] was gaining popularity and software engineers were rethinking the way they developed software. Rapidly changing requirements meant that systems needed to be flexible and able to incorporate change. Defining a data model, or schema in advance is extremely difficult and therefore often never done.

There are four main types of NoSQL databases, each with different specifications, these are: column; key value; graph; and document [205]. Many of the big data performance problems experienced when working with RDBMSs are effectively handled by NoSQL databases. They are extremely efficient when analysing large amounts of unstructured data which is often stored across multiple virtual servers in the cloud [206]. However, relational databases were the product of much research and built using sound mathematical principals. This is not the case with NoSQL databases and often means that consistency is traded for performance and scalability. To support reliability and consistency features, developers are required to implement their own proprietary code. Another big disadvantage of most NoSQL databases is that they are incompatible with SQL, meaning that a proprietary querying language is required, which adds more time and complexity.

MongoDB is a document type of DBMS and used by the MyPathway application in Chapter 6. Document type databases store data in the form of documents and a unique key is used to access these documents. Some of the challenges faced in this study when using data from a NoSQL database are further discussed in Chapter 6.

3.2 Data and process science

Data science is an emerging interdisciplinary field, necessitated by the massive amount of data now available, often referred to as ‘big data’, and the abundance of inexpensive computing power [207]. The data scientist is a hybrid role and requires a skillset in the following domains: computer science; maths; statistics; machine learning; data visualisation; data and process modelling; communication and presentation; and domain expertise. The term ‘big data’ refers to any digital dataset that proves too difficult to store, retrieve and analyse using traditional computing techniques due to its size or complexity [208]. Big data has facilitated advances in many research areas, one of these is the introduction of personalised medicine [209]–[211].

Process science is concerned with uncovering and understanding processes as well as designing interventions for improvement [212]. It is a multi-disciplinary field that combines knowledge from management sciences and information technology [12]. Similar to the data scientist, the process scientist is a hybrid role. The skillset includes the following disciplines: business process management and improvement; process automation and workflow/operations management; stochastics; formal methods; and process mining. Process mining is the bridge between model-based process analysis and data-centric analysis techniques [12].

3.3 Process modelling

As stated above, process modelling is an important skill for the data scientist and has been used in the three studies presented in chapters 5, 6 and 7 of this thesis. In this section the process modelling literature is explored, it specifically focuses on Unified Modelling Language Activity Diagrams, Business Process Model and Notation and Petri nets in order to provide the reader with a better understanding of the process modelling techniques used throughout this thesis.

A process model documents what happens in the development of a process. The level and type of detail included is dependent on the purpose of the model. Process models are created for a wide range of reasons from technical specifications used in system development to visualisations used for communication purposes within a business domain [213]. A process is an ordered collection of activities and may be structured, semi-structured or unstructured [214]. The activities in structured processes are often carried out by following a prescribed order with no room for variation. This type of process often includes operational processes that are repeated. Semi-structured processes are similar to structured, though they allow for some amount of flexibility. Unstructured processes are extremely difficult to describe in advance due to their high degree of variation between instances. Unstructured processes are often highly creative, requiring decisions to be carried out mid-process dependant on certain variables. Healthcare processes are a typical example of unstructured processes, as no two people can be treated identically. The term care pathways is often associated with healthcare processes and are used to document the flow from one activity to another for a patient in a clinical setting [14]. Care pathways may also be used to describe the best practice journey for a patient through an episode of care [215] both locally and nationally as part of clinical guidance. The National Institute for Health and Care Excellence (NICE) clinical guidelines [216] are an excellent example of how care pathways are used nationally to document best practice patient journeys for specific conditions.

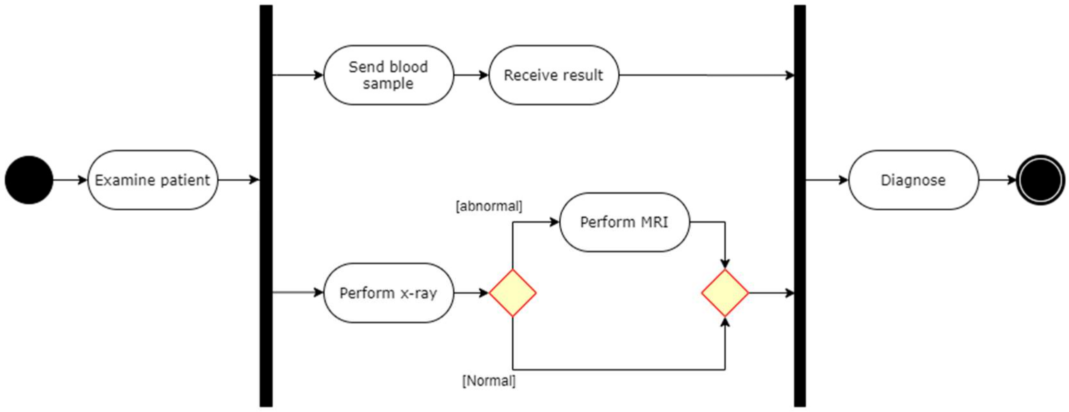
Process models may be produced manually, by applying business process modelling techniques. These models are often developed in collaboration with business domain experts or by referring to the published literature such as the NICE guidelines. Process models may also be derived from data. Throughout this thesis these are referred to as discovered models, though they may also be called data-driven, as-is or learned models. Any model, manually created or discovered may be used as a reference model. Other terms used for reference models include normative, conceptual or theoretical models.

Process models are created for many purposes, some of these are for: discussion and insight; documentation; verification and error detection; performance analysis; animation; specification and configuration; and system design [12]. Many different approaches exist for modelling processes including flowcharts [217], dataflow diagrams [218], Unified Modelling Language (UML) models [219], Petri nets [12] and Business Process Model and Notation (BPMN) diagrams [220]. The process models presented in this thesis have been produced for insight and discussion, verification and performance analysis purposes. Most of these process models have been expressed by using UML Activity Diagrams, BPMN and Petri net models, and the structure and syntax of these will be described in the next three sections.

3.3.1 Unified Modelling Language Activity Diagrams

The Unified Modelling Language is the worldwide standard visual modelling language for the architecture, design and implementation of software systems both structurally and behaviourally [219]. It was first developed by Booch, Rumbaugh and Jacobson in 1995 and later adopted by the Object Management Group (OMG) [221]. UML Activity Diagrams model behavioural control flow information [222]. Control flow is the order in which the activities of a process are executed. Activity Diagrams use a simple, yet powerful notation whereby rounded cornered boxes represent actions, diamonds represent exclusive OR (XOR) decision points and merges, bars represent the splitting or joining of concurrent activities (AND-split or OR-split using guard condition), a solid circle represents the start of the process and an encircled solid circle represents the end of the process. The example in Figure 3.1 models the process for a GP consultation starting with a physical examination of the patient followed by both an X-ray and a blood test. If the results from the X-ray are abnormal an MRI scan is carried out. Once the test results are received and all imaging is complete the GP can then make a diagnosis.

Figure 3.1 UML Activity Diagram example

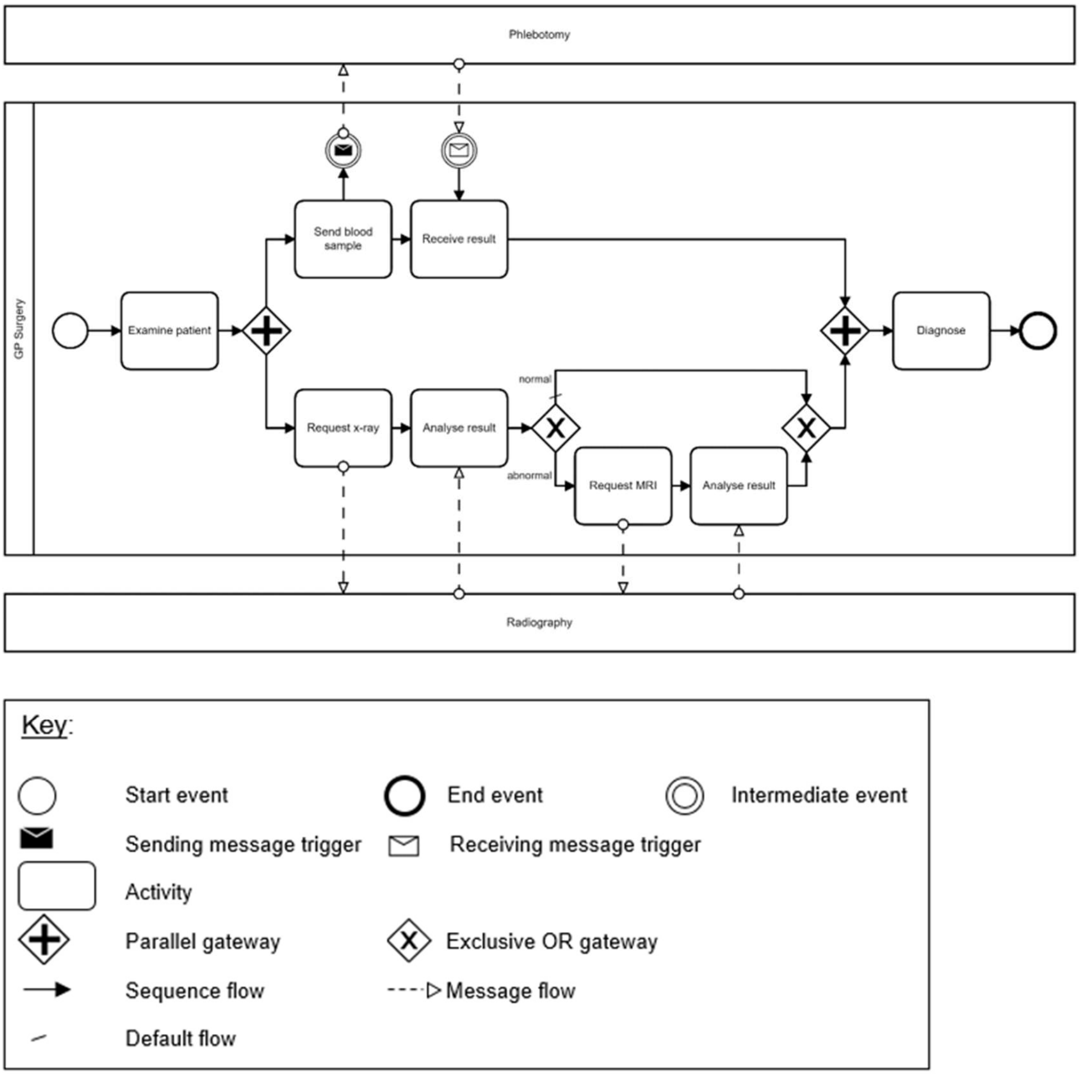


There are many advantages to using UML activity diagrams including that they are designed to be easily understood by both technical and non-technical users. Activity Diagrams offer comprehensive support for the control-flow perspective, though Russell et al. suggest they could benefit from the inclusion of the ‘place’ as with Petri nets (see Section 3.3.3) to capture the notion of ‘waiting state’ [223]. Russell continues that there is no notion of individual cases which could lead to problems when modelling concurrent processes and no ability to model the external environment. However, their suitability for modelling resource-related or organisational aspects of business processes is extremely limited. They are not able to capture many of the natural constructs encountered in business processes such as cases and the notion of interaction with the operational environment in which the process functions. These are limitations that they share with most other business process modelling formalisms and reflect the overwhelming emphasis that has been placed on the control-flow and data perspectives in contemporary modelling notations.

3.3.2 Business Process Model and Notation

The Business Process Model and Notation is a standardised graphical representation for modelling business processes. It was first developed in 2004 by the Business Process Management Initiative. Though is now maintained by the OMG, as the two organisations merged in 2005. BPMN allows known processes, either manual or otherwise, to be modelled as a sequence of well-defined steps. The notation itself is based on traditional flow-charting and therefore familiar to many non-technical users. A process is typically modelled as a sequence of activities that start due to the occurrence of a specific trigger event. The order of activities is determined using sequence flows, which when combined with gateways, define all possible paths through the process. The gateways act as decisions points that use a condition to determine which path should be followed. Message flows are used to model communication with external entities. An example of a BPMN diagram is presented in Figure 3.2.

Figure 3.2 BPMN diagram example



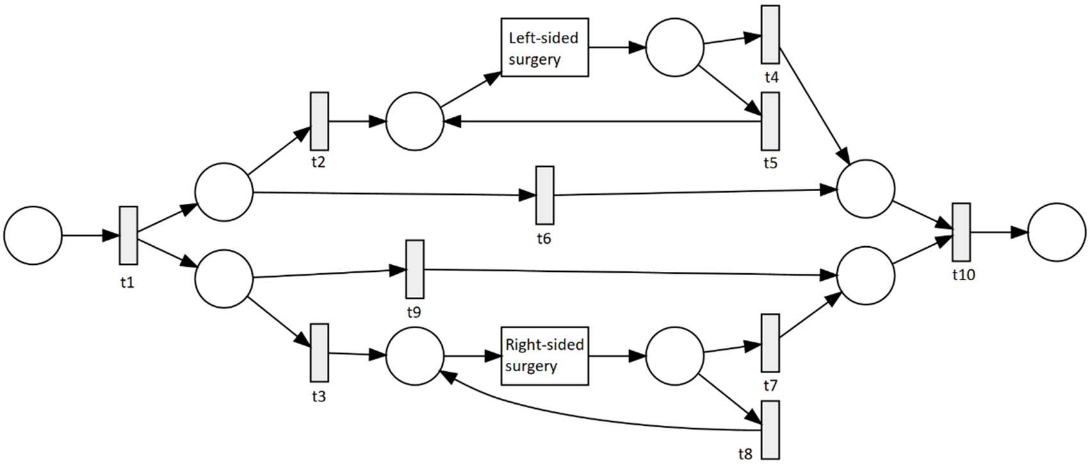
UML activity diagrams and BPMN models can both be used to model the same processes. As they are now both managed by the OMG, it is likely that at in the future the two may merge. A study which performed a comparison between the two, showed that both types of process model could adequately represent the real business processes and both had the same level of readability for all types of user [224].

Over the past 15 years BPMN has become extremely popular for creating process models as minimal training is required in order to understand a basic model. There are a number of software tools available which support the development of models and although BPMN is based on a formal specification, adherence to this standard by supporting tools is varied. Despite its popularity, it has been conceded that 'there is a lot of bad BPMN out there' [225] referring to diagrams that are 'invalid, incomplete, or ambiguous'. This is not the fault of the notation, rather the fault of poor practitioners, often made possible by tools that do not enforce the underlying rules defined within the formal specification. Overall BPMN is a very well-known and well used approach to modelling of business processes. It provides a good mechanism of commination between stake holders and is based on a mature standard. Adherence to this standard however is mixed, and the notation itself is often abused by its own users.

3.3.3 Petri nets

The Petri net modelling technique was invented in August 1939 by Carl Adam Petri at the age of 13 to describe chemical processes. It was first documented in 1962 as part of his dissertation [226]. A Petri net is a directed bipartite graph consisting of two basic node types: transitions, represented by rectangles; and places, represented by circles, see Figure 3.3. Transitions are either visible, meaning they represent an activity within the process or invisible (also known as silent), where their purpose is either to support the structure of the model or to represent a null activity. For example a silent transition may be used to construct an AND-split or to facilitate a loop as seen in Figure 3.3 below.

Figure 3.3 Petri net example showing visible and invisible transitions



Here, all but two transitions are silent and are mainly used to form the structure of the Petri net. *t1* and *t10* form a logical AND split and join, *t5* and *t8* are used to form a loop structure and *t6* and *t9* are used to represent null activities.

The two node types are connected by directed arcs, though no two nodes of the same type may be connected [227]. Petri nets have tokens, shown as black dots, which travel between places to change the state of the model. A transition is enabled, or the activity can occur, if all places directly preceding the transition, known as input places, contain a token. When a transition becomes active, known as firing, it consumes one token from each of its input places and produces one output token for each of its output places. Petri nets have been extended in many different ways to study specific system properties, such as reliability, performance and schedulability [228], [229]. The standard file format for Petri net models is PNML (Petri Net Markup Language) [230].

A disadvantage to modelling using Petri nets is that they can be difficult to analyse as they can rapidly become unmanageable [231] when modelling complex systems. There are over 40 Petri net graphical editing tools available [232], [233] though only a small proportion of these support the PNML standard. Workflow Petri Net Designer (WoPeD) is an open-source Petri net software tool that supports this standard and was created for modelling, simulating and analysing processes [234].

Workflow nets, also known as WF-nets, are a subclass of Petri nets that explicitly model the creation and completion of cases for operational processes. This is done by placing tokens in the source place at the beginning of the instance and the sink place at the end [12].

3.4 Process mining

The term ‘process mining’ was invented by van der Aalst in 2003 and is also often referred to as ‘process discovery’, ‘workflow mining’ or ‘business process management’ [235]. It was first used to discover operational workflow processes [236]. Process mining is now an emerging discipline that bridges the gap between data science (Section 3.2) and process modelling (Section 3.3) [12]. Unlike traditional process modelling approaches it allows for decisions to be made based on the facts from the *actual* data, rather than on assumptions. The goals of process mining are to *discover, monitor* and *improve* real processes by extracting knowledge from event logs, created from data within information systems [12]. An event log is a file containing a collection of events, where an event is a timestamped activity relating to a specific case. An example of an event is: *'01-01-2001 09:05', 'hospital admission', 'P. Smith'*. An event log must contain data at the correct level of granularity, too many events leads to 'spaghetti' type models [12] that are impossible to understand and too few events would result in insufficient information to answer the research questions. The sequence of events for a case is referred to as a *trace*. Events may include additional data attributes, which can be used to extend the model. These data attributes may include resource, location and cost information. Event logs can be used to conduct three different types of process mining which are process discovery, conformance checking and enhancement. Each of these types will be explored in this section. Process mining can be performed from four different perspectives [12]:

- **Control-flow perspective** focuses on the ordering of activities.
- **Organisational perspective** focuses on the information about the resources performing the activities.
- **Time perspective** focuses on the frequency and timing of activities.
- **Case perspective** focuses on the properties of the case, where an example of a case may be an order or a person.

The first published work on process mining was a computer science academic thesis in 1996 by Cook [237] and was originally applied to business processes by Agrawal et al. in 1998 [238] where the term *work flow mining* was used [239]. The number of recent literature reviews on process mining demonstrates the growing interest around the topic [240][241][242][243][244]. In 2012 the IEEE Task Force on Process Mining created the Process Mining Manifesto [245]. This manifesto defines a set of guiding principles and lists a set of important challenges for process mining researchers.

3.4.1 Process discovery

The majority of process mining projects are discovery type projects [59][246][247]. Here the process mining tool takes an event log and applies a discovery algorithm to produce a process model. The most common algorithm used for discovery is the Heuristics Miner [248][241], other ones include the Alpha α algorithm [236], Fuzzy Miner [249] and Inductive Miner [250]. The discovered process models are displayed using an appropriate notation such as activity diagrams, BPMN, Petri-nets and process trees. A process modelling notation must be able to logically represent the event log data in terms of sequence, choice, parallelism and repetition. A discovery algorithm needs to produce process models that are of high quality. The literature suggests a range of methods for the validation of process models, which often includes measuring the model quality in terms of replay fitness, precision, generalisation and simplicity [251].

- **Replay fitness** refers to how well the model can reproduce the observed behaviour from the event log. A fitness score of 100 percent means that all traces in the event log can be replayed on the model from beginning to end.

- **Simplicity** refers to how easy the model is to understand. It is best to generate the simplest model which can describe the behaviour in the event log. There are various ways that the simplicity metric can be calculated including by counting the number of nodes and arcs, or considering the *structuredness* or *homogeneity*.

- **Precision** is the ratio between the amount of behaviour observed in the event log and the amount of behaviour described by the model [252]. Precision is high if the model mostly allows for only behaviour seen in the log. Though care must be taken not to *overfit* the model by allowing it to be too guided by *noise,* or outliers [207]. Algorithms that overfit assume completeness in the log, which is rarely the case. A high precision value indicates the model is less precise. Soundness may also be used to measure precision [253].

- **Generalisation** is the opposite of precision and similarly care must be taken not to allow *underfitting* which is where the model allows for behaviour which is not supported by the log.

Though it is often difficult, a good model is seen as one that scores high in all four of the quality metrics. Three particularly good discovery algorithms, capable of discovering high quality models are the Fuzzy and Heuristics Miners [254], [255], and more recently the Inductive Miner [256].

3.4.2 Conformance checking

Conformance checking is the second most practised type of process mining [257] and evaluate the relation between process models and reality in form of event logs. The model may have been discovered using a discovery algorithm or created manually. Three main techniques exist for conformance checking and these are; token replay, alignments and comparing footprints [12]. Results are shown in both directions, how closely the model reflects reality and how closely reality relates to the model. After conformance checking, it is possible to enhance the process model by repairing it and aligning it more closely with reality or by extending it to add more perspectives such as organisational, time and case [12]. Apart from repairing the model, conformance checking can be done to find undesirable deviations from the model which may suggest inefficiencies or non-compliance within a process. Of the four quality criteria fitness, generalisation, precision and simplicity, replay fitness is the metric most related to conformance.

Conformance checking requires two inputs, these are an event log and a process model. Process models are usually created in the form of Petri nets [258]. Adriansyah et al. in a study looking at the robustness of conformance checking [259] commented on how, without in depth knowledge of the language and algorithm, Petri net-based conformance checking techniques may produce 'false negative' results due to issues with silent transitions (see Section 3.3.3).

Recent progress is being made with new algorithms for the creation of Petri nets for use in conformance checking [260], [261].

3.4.3 Enhancement

The third type of process mining is enhancement. Here the model is extended or improved using data recorded in the event log. This information may include resource, cost and time information in order to analyse time trends, determine bottlenecks and infrequent behaviour or to discover feedback loops [262]. As mentioned above, enhancement techniques can be used to repair a model so it better reflects reality, or they can be used to extend the model. An example of this was done by Badakhshan et al., where they used enhancement techniques to create a performance analysis that helped a hospital emergency room to improve their processes [263].

3.4.4 Process mining methodology

There are various project management methods which have been developed over the past 20 years to help guide researchers through the process mining phases within a project. Many of these methods are based on the more established data mining frameworks such as the Cross-Industry Standard Process for Data Mining (CRISP-DM) [264] and Sample, Explore, Modify, Model, Assess (SEMMA) [265] developed in the late 1990's. The L* Life Cycle Model [245], PM^2 [16] and the Process Diagnostics Method (PDM) [266] are three of the more accepted methods in use today and are described below.

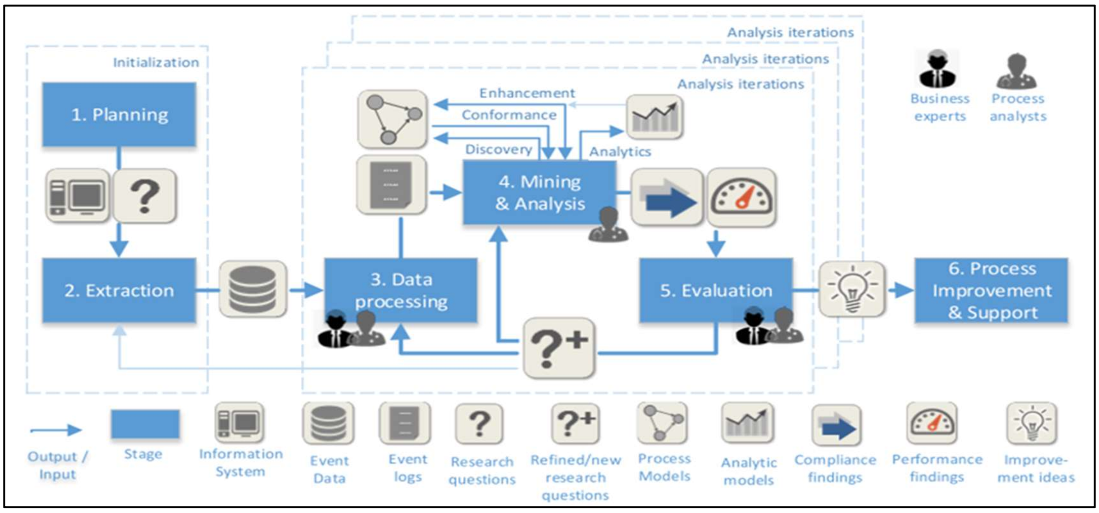
1) L Life Cycle Model*

The L* Life Cycle Model was developed by the IEEE Task Force on Process Mining in 2011 and has five stages. The method begins with a project initiation activity, Stage 0, Plan and Justify. Here the project is defined as belonging to one of three types; data driven, question driven or goal driven. Outputs from this stage are an understanding of the data to be used and of the business domain. During Stage 1 event data, models, objectives and questions are extracted from the source systems and domain experts and event logs are created. Van Eck [267] stated that combining data extraction and event log creation into one stage creates coupling which may result in restricted iterative analysis. Outputs from Stage 1 are used to create a control-flow model and connect the event log in Stage 2. Stage 3 is optional and used to create an integrated process model which can be used for operational support during Stage 4. Syamsiyah [268] commented that Stage 4 is only suitable for structured processes. It aims to detect, predict and recommend appropriate actions based on historical and live, current data which is fed directly to the users without interpretation by the analyst.

2) *PM² Method*

PM² was developed in 2015 by Maikel et al. [16] and covers a wide range of process mining techniques. Unlike the L* method it is suitable for analysing both structured and unstructured processes [268]. It consists of six stages as can be seen in Figure 3.4.

Figure 3.4 A visual representation of the PM² Process Mining Project Methodology [16]



The project initiation phase has two stages. Stage 1 Planning, involves establishing the research questions, identifying the source systems for extraction of event data and forming a multi-disciplinary project team. The inputs are the business processes to be analysed. The scope of the project is defined in Stage 2, event data and optionally process models are extracted from the information systems identified during Stage 1. Domain knowledge is transferred from the experts to the analyst. Unlike the L* method, the extraction of the event data and the log creation and processing are separated into two stages. By separating the stages different views can be created on the same event resulting in multiple event logs [16]. Stages 3, 4 and 5 are intended to be iterative and performed for each experiment or research question. During Stage 3, the event data is transformed into event logs and optionally process models may be used as an input for conformance checking or enhancement. The events may be aggregated to reduce complexity, logs enriched with additional attributes or filtered to reduce complexity. In Stage 4 discovery, conformance and enhancement process mining techniques are applied to the event logs and process analytics may be performed. During Stage 5 the findings from the previous stage are interpreted and results verified against the original data and validated with the domain experts. Here improvement ideas or new research questions may be generated. Finally, in Stage 6 any improvement ideas are implemented and for structured processes, operational activities supported.

3) Process Diagnostics Method

The PDM was developed by Bozkaya et al. in 2009 to enable a quick overview of a business process without the need for domain experts. The method consists of six phases: Phase A Log Preparation; Phase B Log Inspection; Phase C Control Flow Analysis; Phase D Performance Analysis; Phase E Transfer of Results and Phase F Role Analysis. In Phase A the event data is extracted from the source systems and the event log is created. The analyst familiarises themselves with the event log in Phase B and generates some basic statistics to gain further insights before filtering the event log to remove incomplete cases. In Phase C the event log is used to create a control flow model and if a process model is available conformance checking may later be performed. During Phase D, the process models are used to compare throughput times and identify any bottlenecks. Phase E is optional and only performed if resource information is available to create a social network analysis. Finally, in Phase F, the domain expert is presented with the findings to initiate a redesign, adjustments or interventions to the existing business processes.

The PDM is extremely basic and aims to provide a broad overview of the process within a short period of time. As stated by Suriadi et al., for large scale, complex projects PDM is not ideal as it has limited scope and covers only some of the process mining techniques [269]. PDM was discounted for this research programme for these reasons, though predominantly because it emphasises avoiding the use of domain experts during analysis [266]. However, PDM should be considered for less complex projects, where time is restricted. The basic stages in both the L* life-cycle model and the PM^2 process mining project method are very similar. Both methods encourage iterative analysis and were designed to support process mining projects, containing structured and unstructured processes. The aim of these methods is to improve process performance or compliance. However, the L* life-cycle model lacks a detailed technical description, whereas the PM^2 method provides comprehensive guidance at each stage. In this research, a variation on the PM^2 Process Mining Project Method has been used.

3.4.5 Process mining software

There are over 30 process mining tools available [242], [257]. These range from commercial tools to open-source tools used primarily for research. The three process mining tools ProM [270], DISCO [271] and Celonis [272] have been chosen for use in this research programme as they are the most popular and well-established tools, able to fulfil the requirements of this research.

The ProM framework is an open source process mining framework first created in 2005 for use by academics. ProM allows for a wide variety of process mining techniques in the form of *plugins*. The framework is built around the following three concepts: *data objects*, which include event logs and process models; *plugins,* to perform tasks such as event log filtering, process discovery, conformance checking and model enhancement; and *visualisers* for the data objects such as event logs, Petri-nets and BPMN Diagrams [273]. ProM uses the standard formats MXML and XES for importing and exporting of event logs and various diagramming standards, for example PNML for Petri-net models.

DISCO is a commercial process mining tool created by Fluxicon. It was first introduced in 2009, as an easy and fast way for business users to produce process models. Process models are discovered using the DISCO Miner algorithm which is based on the Fuzzy Miner [274]. Discovered process models are 100 percent truthful to the event log data [271] and can be viewed by frequency or duration, and also animated. In addition to discovering process models, DISCO produces process statistics, users can view variants and individual cases and perform a number filtering tasks. DISCO uses the standard formats MXML and XES for importing and exporting of event logs and is therefore compatible with other process mining tools such as the ProM framework.

Celonis was formed in 2011 and is the global leader within the commercial process mining community. Celonis produces uncomplicated models that can accurately reproduce 100 percent of the traces in the event log on the model and was therefore ideal for this research. It produces models by applying the fuzzy miner algorithm [235] to the event log data. An advantage of using Celonis for process mining is that it allows for easy reproduction of the models for other researchers or data scientists wanting to replicate this method with their own data.

In 2003 Mining eXtensible Markup Language MXML [275] was created for storing and exchanging event logs. In 2009 the IEEE Task Force on Process Mining was established and created Extensible Event Stream (XES) [276], which superseded MXML to become the standard exchange format for process mining.

3.5 Process mining in healthcare

To date, ten process mining literature reviews have been carried out in the general healthcare domain. Yang and Su were the first in 2014 [277] to identify 37 publications in process mining for clinical pathways. The articles dated back to 2004 and were classified according to the three dimensions: process discovery; variants analysis and control; and evaluation and improvement. In 2015 Rojas et al. [278] applied a broader scope to provide a general outline of the main approaches previously used, before discussing the use of process mining algorithms, tools and techniques for the analysis of healthcare processes. In 2016 Rojas et al. extended this review to include 74 articles with associated case studies [37]. These articles were analysed by process and data types, research questions, process mining techniques, perspectives, tools and methods, implementation and analysis strategies and medical domain. Shortly after in 2016 Ghasemi and Amyot produced a systematised literature review on process mining in healthcare [279]. They provided an introduction to process mining, before presenting a worked example in the healthcare domain. Three research questions were addressed to provide a comprehensive review regarding the trend, type and quality of health-related process mining research. Finally in 2016, Erdoğan and Tarhan [280] carried out a systematic mapping study to provide an overview of studies using conformance checking techniques within the healthcare domain. They use an example from an intensive care unit to describe the features of a custom process mining tool.

The two most recent general healthcare reviews to date were in 2018. The first by Erdoğan and Tarhan where they updated and extended their systematic mapping of process mining studies from 2016 to find 172 studies in the healthcare domain [281]. The results revealed that process mining in healthcare continues to be a rapidly growing field of research and practice, however there is a lack of studies covering cases from more than one hospital. Batista and Solanas provided a systematic review [257]. After briefly discussing [37], [277]–[279], 55 additional articles were identified and classified according to nine dimensions. The majority of studies used process discovery techniques followed by enhancement, with only six percent using conformance checking. Cancer and emergency medicine were the two most popular medical domains and there was almost an equal divide between studies considering organisational and treatment processes. They concluded that future researchers may consider the challenges of noise, abstraction, standardisation of processes and privacy issues in log files. The final review published in 2021 was by Dallagassa et al. and included 270 articles published before the end of 2020 [282]. This was a systematic mapping study discussing the opportunities and challenges for applying process mining in healthcare. Articles were mentioned in chronological order and grouped by healthcare environment, process mining type, algorithm and area of contribution. The results revealed that there is still difficulty in finding a process model to represent the patient's entire journey, linking healthcare data from primary and secondary care and that also considers patient's outcomes by adding questionnaires. Dallagassa concluded that to successfully utilise process mining techniques in healthcare improvements must be made in the integration of healthcare systems, the ease of collecting healthcare data and the creation of automated processes capable of learning pattern recognition.

In addition to these, some of the more specific literature reviews in process mining within the healthcare domain include: Kurniati et al. in oncology [283]; Kusuma et al. in cardiology [284]; Farid et al. in frail and elderly care [244]; Williams et al. in primary care [285] and Helm et al. in the use of standardised terms in clinical case studies [286]. Other important contributions to this field include the book by Mans et al. on process mining in healthcare [287], the online course by FutureLearn [288] and the Process-Oriented Data Science for Healthcare (PODS4H) workshop in conjunction with the International Conference on Process Mining (ICPM 2020) [289].

More recently, authors from the Process-Oriented Data Science for Healthcare (PODS4H) group identified 10 distinguishing characteristics of healthcare processes [290]. These characteristics are: "D1: Exhibit Substantial Variability; D2: Value the Infrequent Behaviour; D3: Use Guidelines and Protocols; D4: Break the Glass (the need to sometimes deviate from guidelines and protocols); D5: Consider Data at Multiple Abstraction Levels; D6: Involve a Multi-disciplinary Team; D7: Focus on the Patient; D8: Think about White-box Approaches; D9: Generate Sensitive and Low-Quality Data; and D10: Handle Rapid Evolutions and New Paradigms." These characteristics gave rise to 10 types of challenge when performing process mining within the healthcare domain. These challenges are discussed at the end of the following section.

3.5.1 Challenges for process mining in healthcare

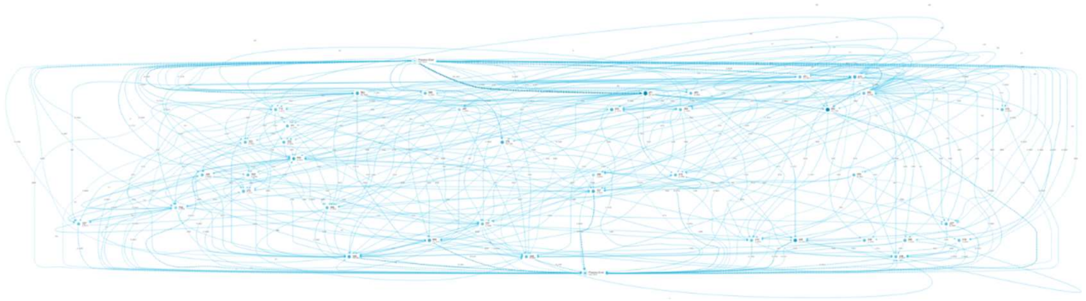
Healthcare is complex, therefore the analysis of healthcare processes is difficult, regardless of the techniques used. Process mining techniques are no exception.

Currently, without a large amount of pre-processing and insight, the results from process mining are often of no use or cannot be understood. There are many reasons why process mining healthcare data is difficult and these are documented as challenges rather than weaknesses in the six sections below. In chapters 5, 6 and 7 references to the challenges are made and ways to approach these challenges are suggested.

3.5.1.1 High variation in healthcare processes

A reason for the recent explosion in process mining projects in healthcare, is that healthcare processes are naturally complex, ad-hoc, dynamic and multidisciplinary, making them difficult to analyse using existing techniques [291]. Van der Aalst often uses the phrase 'spaghetti' type models [12] when referring to healthcare process models due to their appearance, see Figure 3.5.

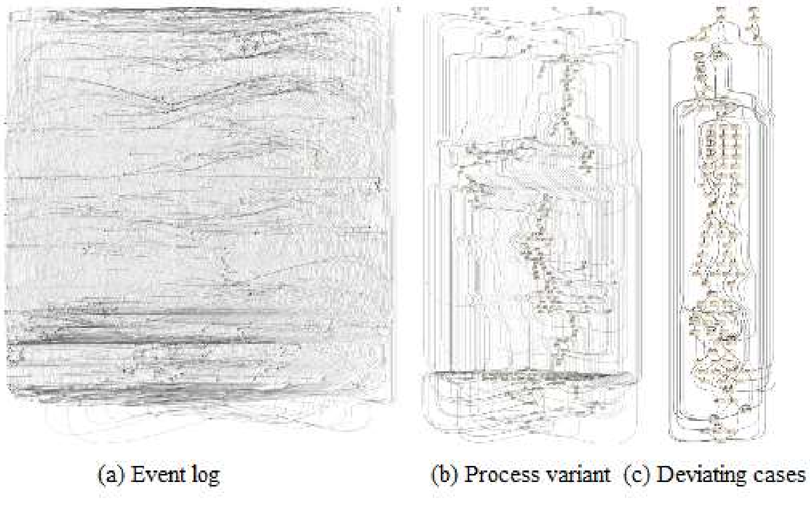
Figure 3.5 'Spaghetti' type process model showing 298,000 cases



Process mining was originally developed to discover workflow processes. These operational processes are relatively straight forward and predictable and have limited variation. As previously mentioned, healthcare is different, there is a high level of variability in the task [290]. There are many reasons for this, one being that no two people are the same. People are individuals and have different heights, weights, socioeconomic statuses, cultures, past experiences, pathologies and co-morbidities. Healthcare decisions are often a result of evidence-based medicine. Evidence-based medicine combines clinical guidelines and protocols with clinical expertise and patient values. However, clinical guidelines are not completely prescriptive and often recommend one of a number of different approaches. Furthermore, clinicians may decide to take a different course of action. This may be due to the availability of resources or patient factors, similar to those mentioned above. In addition to these variables, healthcare is continuously evolving in terms of new clinical knowledge, technological advances and the development of new paradigms [290]. A combination of these factors leads to high variation in healthcare data [10].

Process mining has a range of visualisation tools and uses discovery algorithms to help reduce the amount of variation and noise often present in event log data. Though with healthcare data, in order to understand the output a tremendous amount of insight and pre-processing of the data is required. Pre-processing both in the programming and in the human sense. One way to visualise the data is through trace clustering [292]. Trace clustering enables unstructured processes to be split into homogeneous subsets in order to create a process model for each subset. This may be useful when considering cases that do not fit the norm. An example is presented in Figure 3.6 where Hompes et al. [293] used trace clustering to split unstructured process data into homogeneous subsets, then created a process model for each subset.

Figure 3.6 Trace clustering example



(a) Event log (b) Process variant (c) Deviating cases

This may be a useful technique when working with healthcare data for separating out cases that do not fit the norm.

3.5.1.2 Poor data quality

Healthcare processes tend to be associated with low quality data [26], [290], [294]. This may include missing, imprecise, incorrect or irrelevant data. The primary purpose for the creation of EHR data is for patient treatment and administration, not for research. Healthcare systems are not designed for this secondary use, which is often why the data does not have the same provenance as registry data or data for clinical trials. Healthcare is an intensely human operation, often done by extremely busy people that do not see the information system and data quality as their concern. When entering information, especially when resources are limited and overstretched, it is easy for mistakes to be made. Data is often retrospectively entered, which can lead to incorrect date and time stamps. Clinicians tend to use only a subset of the available clinical codes when entering coded information which may lead to incorrect or imprecise information [290].

The data is only a partial view on reality. There is often a gap between real life and what is recorded in the patient's record. People, processes, organisations and healthcare systems change over time and all these changes impact on data quality. Secondary use of healthcare data for research purposes requires a validated method of data quality assessment. Weiskopf and Weng [295] and Mans et al. [287] propose two similar methods which are discussed in Section 3.6.1.

3.5.1.3 Multiple levels of abstraction

Healthcare data is collected from a wide range of sources including healthcare information systems, sensors, monitors, mobile devises and imaging machines. The data may be recorded in varying levels of granularity from high-level summary information to low-level detailed observations and measures often taken using medical equipment. Often data needs to be aggregated or enriched with additional information in order for it be useful when applying process mining techniques [296]. An important and often difficult step is the pre-processing of data for the event log. This may include the selection of data items, clustering of similar activities or traces, filtering of rare cases and the application of business process techniques [254].

3.5.1.4 Data sensitivity and accessibility

Due to the sensitivity of patient-level healthcare data, legal, ethical and professional considerations need to be taken into account [297] throughout the entire course of the project. Legal considerations relate to a set of rules enforced through the courts, used to regulate society. Ethical considerations are concerned with morality and how we ought to live. Professional considerations relate to a set of guiding principles adopted by a group such as the medical profession, and describe how its members should behave.

When working with healthcare data in the UK, there are many legal constraints and researchers should be familiar with the following: Human Rights Act 1998 [298]; NHS Act 2006 [299]; Health and Social Care Act 2012 [300]; and General Data Protection Regulation 2018 [301]. In order to abide by these rules it is often necessary to ensure that personal data is not identifiable. There are also other considerations such as consent. Ethical approval can be a lengthy and expensive process and should be carried out during the early stages of the project.

Unless there is a specific dataset created for research purposes, knowing how to access healthcare datasets can be challenging. A report by the Medical Research Council [302] offers some guidance. Examples of data custodians and their datasets include the Office for National Statistics for mortality data, Clinical Practice Research Datalink for NHS Primary Care data and Public Health England for the National Cancer Data Repository.

When used to provide end-to-end process models datasets often need to be linked. This is an expensive and time consuming process that is not always possible because of complex, shifting, regulatory and bureaucratic hurdles.

3.5.1.5 Domain expert input

Due to the complexity of healthcare processes it is essential to have the involvement of domain experts throughout the project to help plan the research, guide on medical terminology, language and clinical codes, especially during data extraction, validate outputs, evaluate insights and manage changes. The multi-disciplinary nature of healthcare processes often requires expertise from nurses and other healthcare professionals as well as physicians. Beerepoot et al., [303] describe how nurses were heavily involved in the generation of results. Unfortunately, healthcare professionals are increasingly overworked and often find it difficult to find time for research activities. Therefore time spent with healthcare professionals needs to be carefully planned to ensure maximum benefit is gained.

3.5.1.6 Summary of challenges

As is evident from the literature, process mining in healthcare is difficult and there are many challenges that must be overcome. Many of these challenges have been identified by Munoz-Gama et al., [290] and categorised into the following: "C1: Design Dedicated/Tailored Methodologies and Frameworks; C2: Discover Beyond Discovery; C3: Mind the Concept Drift; C4: Deal with Reality; C5: Do it Yourself; C6: Pay Attention to Data Quality; C7: Take Care of Privacy and Security; C8: Look at the Process through the Patient's Eyes; C9: Complement Healthcare Information Systems with the Process Perspective; and C10: Evolve in Symbiosis with the Developments in the Healthcare Domain." In this article Munoz-Gama concluded that contribution is needed from the process mining for healthcare (PM4H) community to study and suggest new approaches that will address these challenges. During chapters 5 to 8 of this thesis many of these challenges are evidenced and ways to overcome them are presented and discussed.

3.5.2 Process mining in MSK conditions

This section follows on from the review on process mining in healthcare to specifically describe the current state of research on process mining within the MSK domain and to identify opportunities for further research in this area. In the first sub-section, the review process is defined by posing three literature search questions, identifying appropriate sources and by creating a search string using keywords and terms. Results from the review process are presented, before they are discussed in the second sub-section and future opportunities are identified.

3.5.2.1 The review process

To enable a comprehensive search strategy that would set the foundations for this research three literature search questions were formed, these were:

1. Are there any published studies where process mining methods have been applied in the MSK domain?
2. What are the results of previous studies in process mining in the MSK domain?
3. What are the future opportunities for research in process mining in MSK?

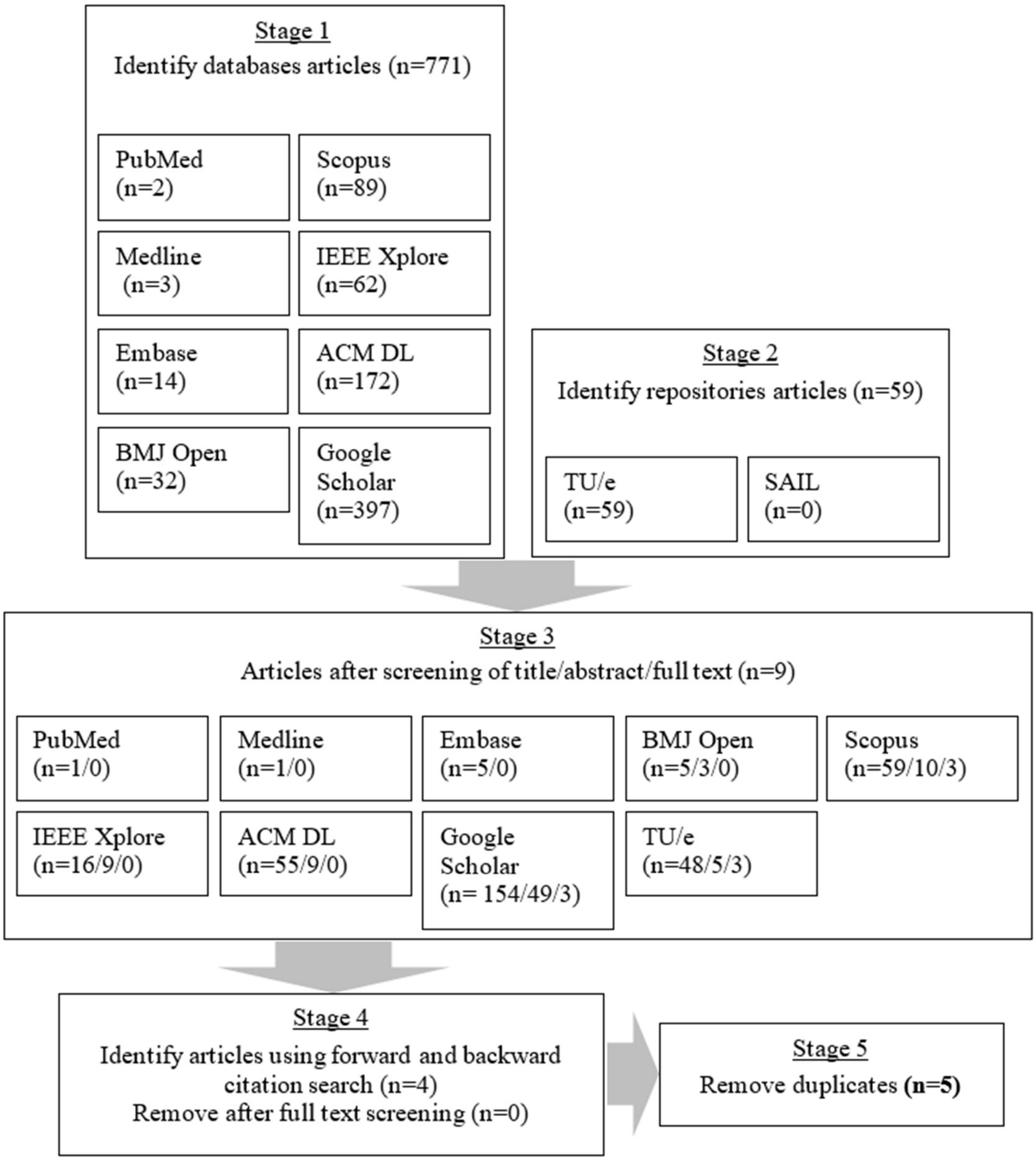
A detailed literature search was carried out for process mining within the MSK domain. A broad coverage was reached by searching the following on-line health, general and computing related research databases: PubMed; Medline; Embase; BMJ Open; and Scopus; Google Scholar; IEEE Xplore and ACM DL. In addition, the two following repositories of publications were searched: process mining at Eindhoven University of Technology [304]; and the SAIL Databank [305]. Once all publications had been screened, a forward and backward citation search was completed for the ones relevant to the field of process mining in MSK diseases. The database selection was based on both recommendations from an article on literature search optimisation for studies relating to MSK disorders [306], and those chosen in similar literature reviews [279], [283], [37].

Within all databases full-text searches were performed with no restriction on date, and where possible only publications written in English were returned. Peer reviewed journal articles and conference papers and Ph.D. and Master theses were included. Any publications that only referenced a publication on process mining in MSK or where MSK was one of many reported on conditions were discounted. An initial search string was guided by keywords and terms frequently used in similar reviews. This was tested and refined before arriving at the final version below:

> *("process mining" OR "workflow mining" OR “event log”) AND (arthritis OR arthrosis OR osteoarthritis OR musculoskeletal OR "musculo-skeletal" OR msk OR physiotherapy OR orthopaedic OR orthopedic OR "back pain" OR "neck pain" OR "knee pain").*

This query was applied to all the health and computing related databases and appropriate medical sub-headings (MeSH) were selected and included where available. Google Scholar was used in Google Chrome incognito mode to avoid potential skewing of results [307]. It was the final database to be searched and returned and returned 854 publications. The review process to that point had proved that the term ‘event log’ had no benefit on the search results and was therefore removed. The review process along with the number of results at each stage can be seen in Figure 3.7.

Figure 3.7 Review process for Process Mining in MSK conditions



Screening for exclusion was performed in three steps: 1) title-based exclusion; 2) abstract-based exclusion; and 3) full text-based exclusion.

The information presented in Figure 3.7 reveals that very little research has been published in the area of process mining within the MSK domain. Stages 1 and 2 present the breakdown of the 830 publications initially identified using the eight research databases and two repositories. Stage 3 displays the breakdown for each source after applying the exclusion criteria, resulting in nine appropriate publications.

Forward and backward citation searching was then performed using these nine, which resulted in no additional publications. Finally, duplicates were removed, giving a total of five publications in the area of process mining in MSK conditions.

Performing this process has answered the first of the three literature search questions. In the following section the results from the nine remaining publications are discussed and future opportunities for research in this field are identified.

3.5.2.2 Discussion and opportunities

The first published work in the field of process mining in MSK was the Master's thesis of Zhou in 2009. Zhou used process mining techniques and the CRISP-DM data mining methodology to analyse data for patients being treated for rheumatoid arthritis at the Máxima Medical Centre in the Netherlands [255]. The ProM framework was used, with the Heuristics Miner algorithm for discovery of the patient careflow and the Linear Temporal Logic Checker and Conformance Checker plugins for conformance checking. The dotted chart and the Performance Analysis with Petri net plugin were both used to gather timing information. Overall, the CRISP-DM method was found to be very useful, despite major differences in the modelling phase between data and process mining projects. When process mining the hospital data the ProM plugins available at the time struggled to handle the high frequency of event types, though this was overcome by creating simplified models. Finally, Zhou produced statistics on the waiting times for new patients to the rheumatology department.

A second Master's thesis was published in 2012 by van Wanrooij, entitled Patient Careflow Discovery [308]. Van Wanrooij used data and process mining techniques to produce a method that would provide insights into the patient careflow. Data was collected from six Dutch hospitals to perform three case studies in order to evaluate the method. These studies are based around patients undergoing surgery for hip replacement, knee replacement and breast cancer. After data cleansing, data mining techniques were applied to the data sets to identify clusters of similar care pathways and to classify main characteristics. Eight activity types were identified and used for process mining, these were: first outpatient appointment; daycare; clinic appointment;

diagnostic activity; surgery; other therapeutic activities; medical imaging; and blood test. Basic control-flow models were discovered using the heuristic miner plugin within the ProM framework and common patterns within these models were identified using the trace alignment plugin. While this work provides an interesting approach, the process mining results lack any kind of useful detail, especially when used outside of the six Dutch hospitals.

In 2019 Valero-Ramon et al. [309] theorised on how applying interactive process mining techniques to data obtained from wearable devises and the Internet of Things (IoT) [310] had the potential to help reduce the number of people with work-related MSK complaints. They provided a proof of concept by creating a simulation to demonstrate how process discovery and clustering algorithms could be used to stratify workers according to their behaviour. The process models were discovered using PMApp, which is a process mining tool created specifically for use in the healthcare domain. Valero-Ramon explained how classic data mining techniques such as Hidden Markov Models [311], Support Vector Machines [312] and Artificial Neural Networks [313] can discover highly accurate models, however, they are seen as black-boxes with no transparency to their logic. In contrast, process mining techniques prioritise understandability over accuracy within the learning process. This allows for bidirectional human-model interaction, where domain experts, such as occupational therapists, can iteratively enhance the models with their own knowledge. Valero-Ramon concludes that process mining techniques can support domain experts in the evaluation and enhancement of process models, as well as using simulation models to evaluate proposed interventions. These evidence-based models can then be used with patients suffering from MSK conditions to suggest beneficial changes to their work-related behaviour.

Process mining techniques were applied by Canjels et al. also in 2019 to knee osteoarthritis patient data from the Maastricht University Medical Centre+ (MUMC) and the outpatient city clinic in the Netherlands with the aim of improving efficiencies in interorganisational patient care [314]. Canjels developed a three-step methodology, where the first step involved the selection and extraction of patient data before applying event aggregation and filtering. The ProM framework was used for the clustering of traces in the second step. Here, four different clustering algorithms were tested, with K-means performing best in relation to average fitness, simplicity of the model, cluster variance and processing time. Five sub-models were discovered using the K-means algorithm. These sub-models, or clusters, contained five different types of similar patient behaviour. The clusters were determined by the type, location and frequency of a patients' visits. For example, one cluster contained patients having only an x-ray followed by an initial hospital consultation. Three of these clusters were labelled as 'non-complex trajectories' and two as 'complex trajectories'. The final step in the method was the visualisation and analysis of these 5 sub-processes. For this, the process mining tool DISCO was used to display the five process models. Results from this study are specific to the MUMC and show that potential efficiencies can be made by ensuring that patients with activities included in the non-complex trajectories undergo treatment at the city clinic rather than the hospital.

Data on patients treated for low back pain was used for a case study in 2020 by Remy et al. [315]. The study discusses the creation of event logs from large, complex data warehouses using standardised vocabularies and domain specific ontologies. The underlying structure of the data warehouse was a star schema [316]. Problems associated with the extraction of event log data from such an architecture were discussed. Questions were identified for the case study around the care pathway and conformance to the clinical guidelines. Process models were generated using the heuristic miner plugin in the ProM framework and conformance checking was carried out manually by cross-checking the findings to the clinical guidelines. The discovered models showed the most common sequence of events to be diagnosis followed by non-pharmacological treatments or prescription painkillers. Remy stated that the results must be interpreted with caution as multiple visits could belong together and therefore the initial diagnosis could be stated in a previous visit. Whilst this article reported on some interesting ideas, low back pain can be a chronic condition and a more in-depth discussion on the method used to correlate events specifically for patients with chronic conditions would have been welcome. Also to add confidence to the results, it is not sufficient to simply consider the visits that exist in the system.

Consideration must also be given to whether a patient was first diagnosed prior to the EHR, or at a hospital or primary care facility that used a different health system.

According to the results of this literature search, no studies to date have published work demonstrating how process mining techniques can be used to produce logical pathway models for patients diagnosed with MSK conditions. A challenge that is foreseen would be the modelling of sidedness. In order to produce accurate process models, showing common pathways, it is often important to know whether the diagnosis, appointment, test, procedure or treatment is related to the left or right side of the body.

3.6 Evaluation techniques

Section 3.6.1 will discuss methods used to assess the quality of data used to create process models. Section 3.6.2 will discuss verification techniques that can be used to evaluate process models. Section 3.6.3 will discuss validation techniques that may be carried out to evaluate process models, their associated data, such as duration, and any results, findings, insights or conclusions generated from these models which are appropriate for consideration within this research. For ease of reading, results, findings, insights and conclusions shall be collectively referred to as results. If necessary, it will be made clear where a technique is only relevant to either a hand-drawn model or a discovered model.

Evaluation may be performed for a variety of reasons. The focus may be on the understandability, usability or the quality of a model, it may be to confirm or assess the relevance, clinical plausibility or generalisability of a result or it may be to check for conformance [317].

3.6.1 Validation of data quality

When creating process models from data, the quality of the results is directly dependant on the input data, i.e., Garbage-In Garbage-Out (GIGO). Therefore, it is important to assess the quality of the data being used to provide the analyst with insights into problems that may arise and negatively impact on the reliability of any results.

Mans et al. in their book on Process Mining in Healthcare [287] identified 27 data quality issues that may compromise the validity of results. They suggested a Data Quality Matrix where data issues are categorised into four types: missing; incorrect, imprecise and irrelevant. This matrix is used in Section 6.3.5.1 to perform a detailed evaluation of the final data extract.

Weiskopf and Weng [295] reviewed the methods and dimensions of EHR data quality assessment. They created a framework considering five dimensions, these are: completeness, correctness, concordance, plausibility and currency. A helpful rating system for data quality was suggested by the Process Mining Manifesto which ranges from one to five star [245]. Kurniati used the Weiskopf and Weng framework as a generic data quality assessment approach for her analysis on the impact of changes in user interfaces of EHR systems on clinical pathways.

3.6.2 Process model verification

Verification is concerned with measuring the quality and correctness of a model in terms of syntax and structure [318]. The various dimensions of quality include correctness, completeness, conciseness and consistency. Verification is carried out to ensure that the model is properly designed and no errors exist. For hand-drawn models this is usually a case of checking for errors made by the modeller or for errors in the implementation of the notation. The latter is less likely if a CASE tool is used rather than a drawing package, as CASE tools generally enforce the underlying rules of the notation. For discovered models, verification is concerned with checking for errors caused by the algorithms.

Soundness [319] is a technique used to verify the quality of a Petri net. A Petri net is declared as being sound if the following properties are true [12]:

- **Safeness**: places cannot hold multiple tokens at one time (boundedness)
- **Correct completion**: if the sink place (end) is marked, all others are empty

- **Ability to complete**: always possible to reach the final marking
- **No dead parts**: all transitions are able to be fired (liveness)

Various tools allow for soundness checking, two of these tools include Woflan [320] and WoPeD [234]. Soundness can easily be adapted to measure the quality of different kinds of process models [12].

Others [321], [322] have used anti-patterns to identify anomalies in BPMN models and WF-nets. These are patterns that appear to provide a solution, though in reality are incorrect [323]. The goal is to describe repeated errors to enable them to be identified and repaired. Suchenia and Ligęza [318] discuss other techniques and model checking tools available for the verification of process models.

3.6.3 Validation of models and results

Once process models have been verified to check for internal correctness, validation should be carried out to check the semantics and assess the quantitative and qualitative correspondence with reality [324] or with other models and results. The goal is to make the models useful, so that they address the correct problem and provide accurate information. Validation is performed by systematically comparing the model outputs and results to independent observations. These observations may be gathered using a range of different validation techniques which are discussed in the sections below.

3.6.3.1 Conformance checking and cross-validation

Conformance checking, as discussed in Section 3.4.2, is a technique used to check that the model aligns with reality. Algorithms are used to check if the behaviour present in the event log follows the rules specified in the model [325] and vice versa. When analysing trajectories of patients with sepsis [326] Mannhardt and Blinde applied conformance checking techniques using the ProM Multi-perspective Process Explorer (MPE) plugin (see Section 4.1.3.3) to validate their hand-drawn model. When the results were presented to a domain expert using the MPE, they commented that process mining was a magnificent way to understand the patient flow.

When process models are generated from the data, it may be beneficial to use validation techniques from data mining and statistics such as the holdout method or cross-validation [12], [327] to ensure that the model is not over-fitting to the data. The term 'internal validation' is often used to describe the process where the same dataset is used to create and validate the model. With the holdout method, the data is divided so that part can be used to create the model and part can be used to test it. This is fine with big data, though when the amount of data is limited stratified cross-validation techniques are often used [328]. Van der Aalst in the Process Mining Manifesto [11] recommends k-fold cross-validation for use with process models. This has been done successfully by Kusuma et al. when implemented to validate disease trajectory models [329]. Here the data is split into k equal parts, often with the use of stratified sampling. K−1 parts can be used to learn the model, and conformance checking techniques can be used to assess the result with respect to the remaining part. This is repeated k times. Cross-validation methods make use of the entire dataset for model creation and provide insights and confidence toward its quality. Kurniati et al. [24] also used conformance checking techniques to assess the quality of oncology process models. The metrics generated showed a high degree of fitness and precision, indicating that the discovered model allowed for the behaviour observed in the event log but did not allow for too much unrelated behaviour. As the models had been discovered from data, they repeated the evaluation steps five times using k-fold cross-validation to ensure the model was not over-fitting to the data. The five results were averaged to produce two single metrics.

When models are to be used with previously unseen data or for prediction purposes, holdout or cross-validation methods [12] may be employed. However, if the model's purpose is to visualise the data from a single dataset or the model is to be used for data cleansing purposes, then this step may not be necessary [330]. In [12] van der Aalst comments that a problem with cross-validation is that you are always limited by the range of examples in the event log which may not cover all situations. This is true when performing any form of conformance checking and one reason why other methods of evaluation, such as using domain experts is important.

3.6.3.2 Validation of models using different datasets

Though k-fold cross-validation is considered the industry standard for validating discovered or learned models, it may not be sufficiently rigorous in critical settings such as healthcare [331]. In such domains subtle differences within the dataset are often important and may render the model unusable in a different setting, for example, a hospital or a city [332]. Validation using independent data from that used to create the model is known as external validation. External validation is important to establish the generalisability of the model [333]. Ho et al. recommends external validation in critical settings as internal validation techniques are naturally highly heterogeneous and tuneable. In addition to this, if the original dataset is biased, then the validation is also biased [333].

Regardless of whether the model is discovered from data or created by hand, using data from more than one geographic region or demographic will better ensure the generalisability and usefulness of the model.

3.6.3.3 Validation using statistical techniques

Statistical techniques can be used for either hypothesis generation or for hypothesis testing. Statistical inference is a way of validating results in wider populations to see if they are reproducible and fit with previous hypotheses and evidence published in the literature. The numerical results generated from previous studies, known as hypotheses, are tested against your results to see if there is a significant difference. However, this can only be done if previous results are available and are reported in sufficient detail. Different tests are carried out depending on the type of data being compared. When results are presented as percentages, such as sex, test of proportions [334] are often used. Data following a symmetrical, normal distribution, such as height and weight, should be summarised using a mean. For the testing of mean values t-tests [335] are often used. Data that has a skewed distribution should be summarised using a median value. Waiting times often produce skewed data as they often contain outliers. For the comparison of median values Wilcoxon tests [334] may be used.

If no prior hypotheses are available, then statistical techniques may be used to generate them and to identify what is potentially clinically important. This is known as descriptive statistics [336] and has been used in Section 6.4.1 to describe the event log data. With descriptive statistics, often the data is characterised using counts, percentages, means and standard deviations, medians and first and third quartiles and modes.

3.6.3.4 Validation using domain experts

A challenge when analysing process models and the results generated from them is that analysts are usually not domain experts and have difficulty determining the cause of unusual or unexpected results [16]. Therefore, a technique often carried out is discussion with domain experts [246], [337]. It must be noted, that when validating models created with the help of domain experts, different ones should be used for the validation. Quantitative and qualitative feedback may be collected, often using questionnaires, during structured or semi-structured interviews, surveys, walkthroughs and focus group sessions. Information gathered using these techniques should relate to the level of understandability, usability and the correctness of the models and the experts should interpret and explain results as well as determine their relevance.

Koorn et al. in their recent literature review for the evaluation of process mining findings [317] identified a need for a more systematic approach for qualitative evaluation where domain experts are involved. They proposed six validation strategies should be considered including: 1) Engagement and understanding of the field; 2) Triangulation, this is when multiple data sources are used to study the research problem; 3) Peer review or external audit; 4) Refine work hypothesis, here cyclical hypothesis refinement is carried out for each negative case analysis; 5) Clarify biases, reflects on possible biases; and 6) Member checking, which involves interviewing participants, analysing the data, then confirming the credibility of the findings by reviewing them with the interviewees.

Alvarez et al. used domain experts to validate the results from a study that was carried out to understand role interactions in the Emergency Room [246]. They prepared a series of five open and closed questions to be delivered during an interview session via a questionnaire. At the beginning of the interview, the models, any assumptions and the findings were presented to the experts. Questionnaires were also used by Lira et al. [338] to collect feedback from student medics after process mining techniques were used to improve the effectiveness of the training.

In [339], a goal-driven evaluation method based on process mining in healthcare was introduced Erdogan et al. They describe the GQFI Table, which was used during the initial stage of their study to address the project goals (G) and research questions (Q). Process mining features (F) where identified that could be used to help measure key performance indicators (I). During the analysis, values for the KPIs were entered and used along with the answers to the questions as input during the evaluation stage. This table has been amended and used during a structured interview in Chapter 7.

3.6.3.5 Validation using the published literature

Models and results can be validated by comparison against the published literature. Examples may include clinical guidelines such as NICE or published statistics, such as the average duration between knee replacement surgeries in the National Joint Registry (NJR). Models and results can be cross-referenced to others that have been produced using similar or different methods. Benevento et al. [340] performed qualitative analysis in order to validate the accuracy of their models. They used Interactive Process Discovery (IPD) [341] to model four healthcare processes. Each model was examined semantically and compared against a set of medical rules extracted from the clinical guidelines. Results from the study demonstrated that IPD provides a more understandable, accurate and guideline compliant model compared to existing techniques. However, it must be remembered that the recommended clinical guidelines represent the should-be model, which may differ from the actual process as defined in the event log. Fox, discussed assessing compliance with care pathways and clinical guidelines in his thesis on applying process-oriented data science to dentistry [342].

3.6.3.6 Validation using direct observation

Before the introduction of EHRs, observational research was often carried out to produce the as-is process model [343]. Similar to deriving process models using data, direct observation provides a way to understand the process as it happened, rather than as it should have happened [344]. Observational studies are far more expensive and time-consuming when compared to data and process mining methods that use routinely collected EHR data. However, observation is a technique that can be used for the validation of process models. There are similarities to conformance checking, though rather than checking the model against historical data in the event log, it is checked by observations in real time. Additionally direct observation can help to understand certain complexities that may otherwise be poorly understood [345]. Unlike using historical data, it is doubtful that all cases will be observed, therefore a sample of the behaviour must be selected. To do this, in advance some concept of the incidence of observable events needs to be understood. Catchpole et al. created a framework for direct observation. In [346] they present the observation process and discuss the different stages, beginning with sampling and include recording, coding and analysis. They conclude that, though often overlooked, the progress from the ‘work as imagined’ to understanding the ‘work as done’ is key when developing an observational approach.

3.6.3.7 Summary of validation techniques

In practice, depending on individual circumstances, a combination of the techniques described above will be used for validation. Circumstances may include: availability of domain experts; access to the problem domain, where the original domain cannot be accessed, one similar may suffice; and the novelty of the research. Where research is completely original, it is likely that no similar studies will be available for the comparison of results.

3.7 Background summary

Within this research process mining techniques have been applied to healthcare data to help answer the primary research questions posed in Chapter 1. To provide a sufficient background for the rest of this thesis and to assess the current state research on process mining in MSK the related healthcare and technical literature has been explored and discussed in chapters 2 and 3. Chapter 2 examines the literature on healthcare systems, clinical coding standards and some of the different ways in how health change can be measured. It concentrates on the three methods used in this research which were disease trajectories, patient reported outcome measures and care pathways. The final part of the healthcare background reviews the literature to provide a background for the four common MSK conditions used in this thesis; osteoarthritis, knee pain, back pain and gout.

Chapter 3 completes the background for this thesis by exploring the technical literature including the emerging field of data science, from where process mining resides. The three process modelling techniques used in this thesis have been introduced, these are activity diagrams, BPMN and Petri-nets. The discipline of Process Mining was examined, resulting in identification of the three sub-disciplines of discovery, conformance checking, and enhancement. This was supported by identification of the main methodologies applied to process mining based projects. Process mining software tools were introduced before considering some of the known challenges faced when applying process mining techniques to healthcare data. Finally, different methods of evaluation were discussed. This included verification techniques for process models and validation techniques for process models and the results generated from these models.

There are many challenges when process mining healthcare data. By reviewing the current literature this chapter has provided insights into some of these challenges which will be further explored throughout this research.

Chapter 4
Methodology

In Chapter 2, a range of healthcare problems and challenges are discussed. In Chapter 3, some of the technologies that can be used to address these problems are reviewed and discussed. In this chapter, the methods to be applied to the three datasets, in order to better understand these problems and help to answer the primary research questions are presented. Different techniques have been applied to the datasets and discussed for each study.

Section 4.1 describes how the method was used at each stage of the three studies. Section 4.2 builds on the introduction given in Section 1.4 to provide a full description of the three datasets used for these studies.

4.1 Method

Three popular process mining methodologies were referred to in Section 3.4.4, these were the L* Life Cycle Model, PM^2 and the Process Diagnostics Method. The method used for the three studies within this thesis was adapted from the Process Mining Project Methodology PM^2, as its iterative approach fitted well and the detailed steps helped to guide the research. For this research, stages 2 and 3 have been combined allowing for multiple iterations of ETL. There are two reasons for this: 1) a multi-stage extraction processes enforced by the data provider (ADI) in Chapter 6 and 2) direct access was available to entire datasets in chapters 5 and 7. This eliminated any requirement for an initial bulk extract and allowed for a more experimental and iterative approach to be taken.

Table 4.1 shows the differences between the stages of PM^2 and the method used for this research. Stages 2 and 3 from PM^2 were combined into Stage 2 ETL and Stage 6 from PM^2 has no support element. Stages highlighted in blue indicate that the stage is iterative.

Table 4.1 Stages of the method, adapted from PM^2

PM^2	This research
1. Planning	1. Planning
2. Extraction	2. Extract, Transform and Load (ETL)
3. Data processing	
4. Mining and analysis	3. Mining and analysis
5. Evaluation	4. Evaluation
6. Process improvement and support	5. Process improvement

4.1.1 Stage 1: Planning

As described in Table 4.1, for this research each study began with Stage 1, Planning. Planning followed the steps from the PM^2 method which consists of constructing a project team, selecting the business processes and identifying the primary research questions.

The project team consisted of a carefully assembled supervisory team from the University of Leeds. Alongside the author, who adopted the role of project lead and data scientist, the team included one other data scientist, two clinical experts, an epidemiologist and a statistician. The data scientist being Mr Owen Johnson, a Senior Fellow from the School of Computing. Professor Philip Conaghan, a Professor of Musculoskeletal Medicine and Dr Sarah Kingsbury, an Associate Professor and Musculoskeletal Strategic Lead, both from the Leeds Institute of Musculoskeletal Medicine (LIRMM) adopted the role of clinical expert. Dr Mar Pujades Rodriguez, a senior clinical epidemiologist from the Leeds Institute of Biomedical and Clinical Sciences and Professor Paul Baxter, a Professor of Biostatistics & Education within the Leeds Institute of Cardiovascular and Metabolic Medicine (LICAMM) completed the project team. Team members occasionally adopted transitory roles such as data owners. Additional resources were recruited, such as business and system experts for the individual studies and these are introduced in the appropriate section.

The business processes analysed during this research were based on processes for patients diagnosed with MSK conditions, where specific conditions included gout, knee and back pain, including osteoarthritis. Three different cohorts of patients were chosen, one for each study. The business process for Chapter 5 was disease trajectories, for Chapter 6 was physiotherapy patient care pathways, and for Chapter 7 was knee pain surgery pathways. Selecting such a diverse range of business processes was key to experiencing a variety of different opportunities and challenges within the application of process mining in healthcare. The method deviated slightly from PM^2, as the majority of the data quality assessment was deferred until after data extraction in Stage 2.

After careful collaboration with the clinical experts, the final part of the planning stage was to identify the primary research questions, which would either prove or disprove the key hypothesis. During the project initialisation phase primary research questions were defined (see Section 1.3). These questions were refined to provide study-specific research questions, which would be answered using the datasets for each of the studies presented in chapters 5, 6 and 7.

4.1.2 Stage 2: Extract, transform and load

For this research Stage 2 followed the basic steps included in Stage 2, Extraction and Stage 3, Data processing, of the standard PM^2 method.

The **Extraction** stage of the PM^2 method consists of the following three activities: 1) determine scope; 2) transfer of process knowledge; and 3) extraction of event data. The extraction scope for each of the three studies was based on the analysis required to answer the study-specific research questions. This required an understanding of each dataset to determine the data attributes, level of granularity and study period necessary for the analysis. In addition to the three data items required for an event log which are ‘case’, ‘event’ and ‘timestamp’, other case or event attributes were often extracted for use during the data transformation stage. In parallel with data extraction, process knowledge was transferred, which often included the collection of existing process models. Dotted charts [347] may be used during the extraction stage to provide a high-level visualisation of the extracted data. Viewing the data in this way can help to verify the semantics of the extraction code, assess data quality and develop hypotheses about the data which can be later tested.

The **Transformation** stage of this research programme followed the Data processing stage of the PM^2 method and consisted of the following activities: 1) creating views; 2) aggregating events; 3) enriching logs; and 4) filtering logs. Before carrying out these process mining activities it may be helpful to document complex transformations rules using a Unified Modelling Language (UML) class model [348].

1) Creating views

Event logs are created in order to generate a specific view on the data. The case classification, also known as the case notion, will determine this view by combining all event instances related to a particular case. For example, there is a big difference between a 'Patient' and a 'Patient referral'. In the case of 'Patient', all events related to that patient would be included to form a process instance. This could include everything over the patients' life-time. Whereas 'Patient referral' would only include events related to a particular referral. Event classes must also be identified, these are the activities that will be included in the process instance. The notion of a case in this thesis was different in each of the three studies. In Chapter 5 a case using the MIMIC-III data was defined by a disease ordered pair, in Chapter 6, using the MyPathway data, a case was defined by a patient referral and in Chapter 7, a case using the SAIL data was defined by a patient.

2) Aggregating events

It is often necessary to aggregate events prior to mining to reduce unnecessary complexity within the process models. An example of aggregation used in this research, is in the creation of disease trajectory models in Chapter 5. When discovering disease trajectories using ordered pairs of diagnoses, only high-level diagnoses at ICD-9 level 3 were included. A second example is in the creation of patient pathways with the MyPathway data in Chapter 6. When attending an outpatient visit, all patients are tracked at various points through the hospital. This level of detail was unnecessary, therefore events were merged to create a single 'attended outpatient' event. A final example is given for with the SAIL data in Chapter 7. All surgery events are recorded with their low-level OPCS-4 procedure codes. In this study, primary total knee replacement (TKR) surgery events are made up of 13 low-level codes.

3) Enriching logs

Event logs may be enriched with additional attributes to make process models more meaningful. This can be done by deriving additional events from the extracted data or by adding external data to the event log. In this research, logs were enriched in Chapter 6 by deriving new event classes. For example, it is important to know whether an outpatient appointment is new or a follow-up appointment. In order to create these two new event classes, triggering event data along with relative positional information from within the pathway is used. A second example is given using the SAIL data in Chapter 7. When creating a knee pain surgery pathway it is crucial to know whether the surgery has been performed on the left or the right knee. All low-level surgery events in the EHR are recorded without laterality. Laterality is recorded as a separate event on the same day. Therefore new event instances must be created by combining the two events.

4) Filtering logs

Filtering is also performed on event data to reduce complexity. There are three types of filtering, these are based on attributes, variants (clusters) or compliance. In this research, an example of attribute filtering can be seen with the SAIL data in Chapter 7. Here, only patients permanently resident in Wales were included in the analysis. An example of filtering using variants provided using the MyPathway data in Chapter 6. Patients were clustered depending on the body for which they were referred. Due to data volumes, all patients not referred for knee or spinal pain were not included in the event log.

For this research, the **Loading** stage was performed by importing the event log files into the three process mining tools, including DISCO, ProM and Celonis. The minimum data items required by any process mining tool are ‘caseID’, ‘activity’ and ‘timestamp’. After selecting the CSV event log file, each column was assigned one of the three data item types above and if required, a timestamp format is defined. The file encoding is selected, in all cases this was ‘UTF-8’ and whether or not quotes are used in the file. The process mining tool DISCO was used to create disease trajectory models using MIMIC-III data in Chapter 5, it was also used to perform a single task with SAIL data in Chapter 7. This was to select and export a list of patients when investigating conformance checking violations with the clinical experts. The process mining tool Celonis was used for process discovery using the MyPathway and SAIL data in Chapters 6 and 7 respectively. Finally, the ProM framework was used with the SAIL data in Chapter 7 for conformance checking of the data against a reference model and for enhancement of both the model and the data.

4.1.3 Stage 3: Mining and analysis

During this section, data and model analytics are created before performing one or more of the different types of process mining. During the following four sub-sections the four types of mining and analysis: process analytics; process discovery; conformance checking; and enhancement are discussed.

4.1.3.1 Process analytics

Process analytics use complementary techniques to process mining in order to provide further insights into the data. They often originate from data mining, visual analytics or statistics. Outputs may include process trees, charts and graphs, visualisation of traces and dotted charts. Process analytics are generated to help characterise the data and enhance the process models by providing more meaningful measures and visualisations.

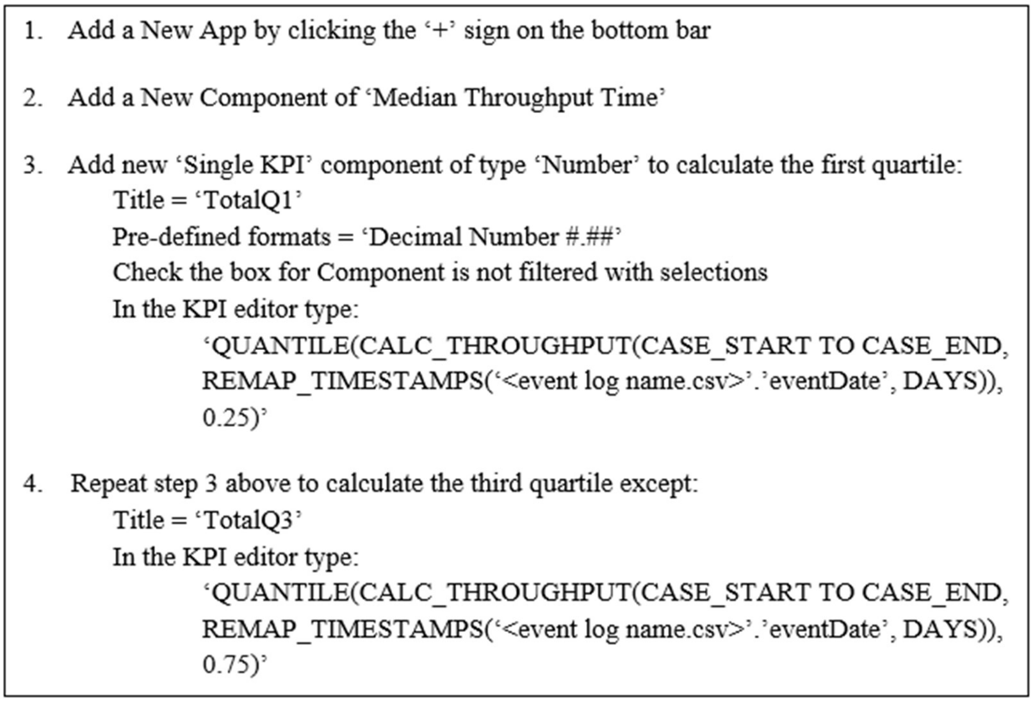
1) Data characterisation

The R Studio version 1.1.442 [349] and Microsoft Excel 2016 [350] were used to produce additional profile information on the MyPathway datasets in Chapter 6. This allowed for informed inferences and conclusions to be drawn from the results. For each event log, or patient cohort, a matching CSV file was created, containing data at the patient referral level (rather than at event level) for the calculation of statistics. The information extracted included the total number of patients based on body part referral (e.g. spine, knee), variable names and data types, summary statistics by age, sex and health outcome. Boxplots, charts and graphs were generated using both R Studio and Microsoft Excel to present this information.

2) *Additional process model measures*

The analysis of event log data in chapters 6 and 7 required additional measures to the standard ones in the Celonis software. These measures included the lower and upper quartile range for both the total case durations and the durations between events. To calculate this information two custom Key Performance Indicators (KPI) were created. Custom KPI 1 calculates the total case duration by performing the steps described in Figure 4.1.

Figure 4.1 Steps to create inter-quartile range for total throughput time in Celonis (Custom KPI 1)



1. Add a New App by clicking the '+' sign on the bottom bar

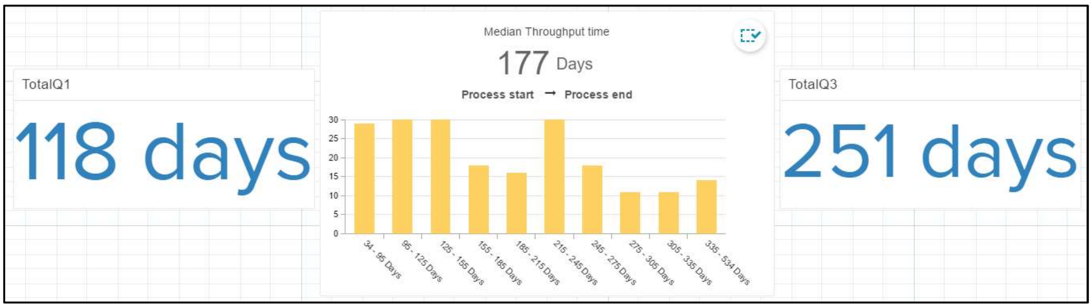
2. Add a New Component of 'Median Throughput Time'

3. Add new 'Single KPI' component of type 'Number' to calculate the first quartile:
   Title = 'TotalQ1'
   Pre-defined formats = 'Decimal Number #.##'
   Check the box for Component is not filtered with selections
   In the KPI editor type:
   'QUANTILE(CALC_THROUGHPUT(CASE_START TO CASE_END, REMAP_TIMESTAMPS('<event log name.csv>'.'eventDate', DAYS)), 0.25)'

4. Repeat step 3 above to calculate the third quartile except:
   Title = 'TotalQ3'
   In the KPI editor type:
   'QUANTILE(CALC_THROUGHPUT(CASE_START TO CASE_END, REMAP_TIMESTAMPS('<event log name.csv>'.'eventDate', DAYS)), 0.75)'

Figure 4.2 presents the output after performing the above steps.

Figure 4.2 Results from Custom KPI 1 showing the total throughput time



TotalQ1

118 days

Median Throughput time

177 Days

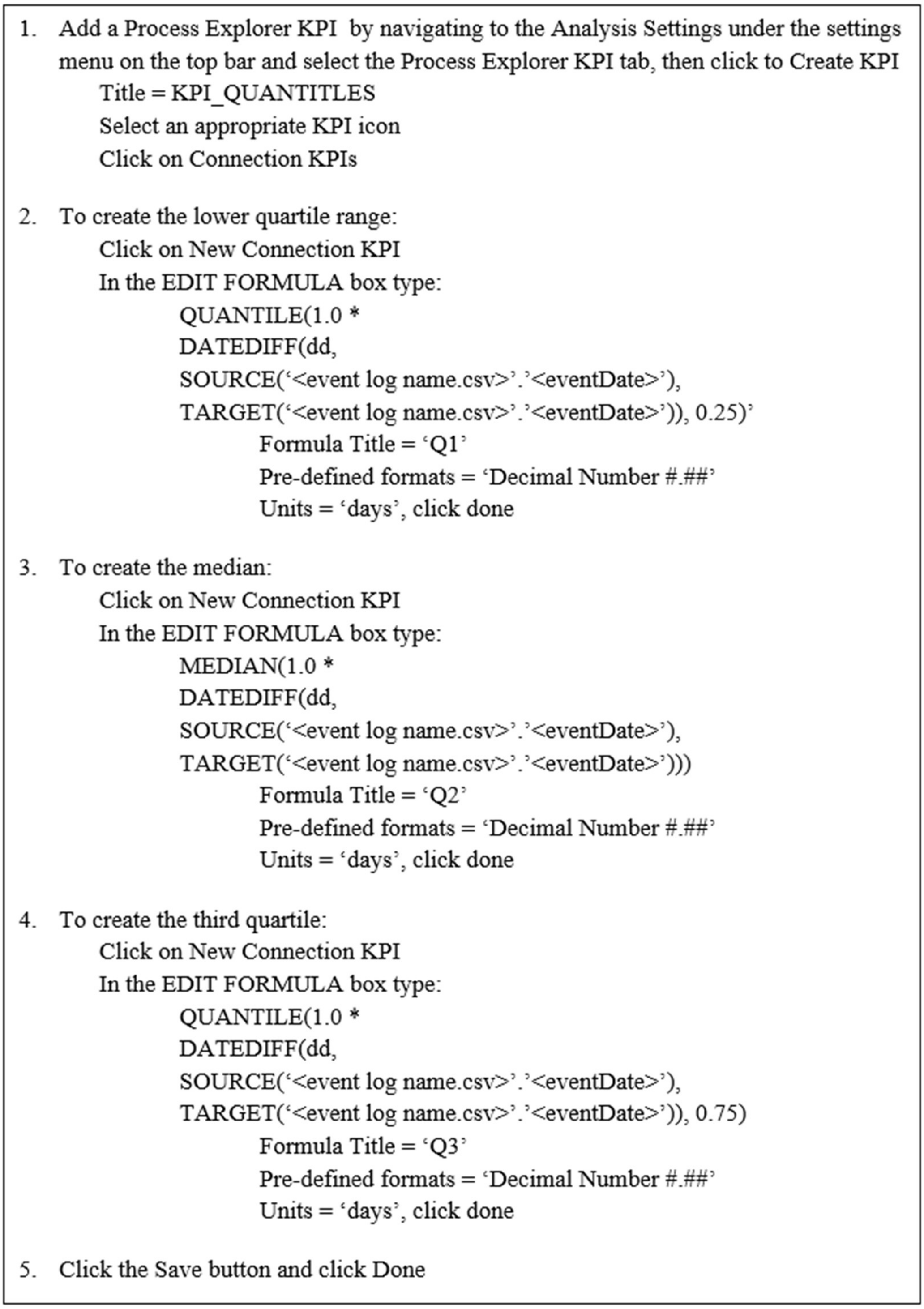
Process start → Process end

TotalQ3

251 days

Figure 4.3 describes the steps required to create Custom KPI 2 which calculates the time intervals between activities.

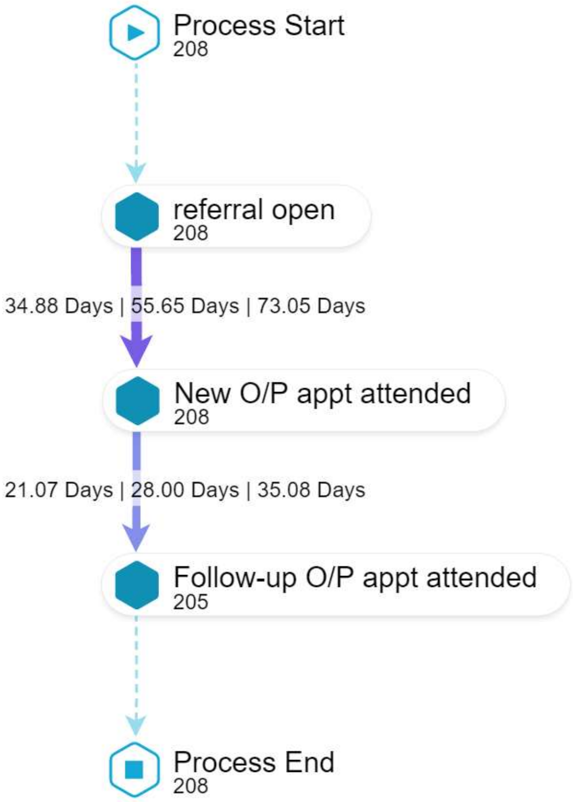
Figure 4.3 Steps to create time interval between activities in Celonis (Custom KPI 2)



```
1.  Add a Process Explorer KPI  by navigating to the Analysis Settings under the settings
    menu on the top bar and select the Process Explorer KPI tab, then click to Create KPI
        Title = KPI_QUANTITLES
        Select an appropriate KPI icon
        Click on Connection KPIs

2.  To create the lower quartile range:
        Click on New Connection KPI
        In the EDIT FORMULA box type:
                QUANTILE(1.0 *
                DATEDIFF(dd,
                SOURCE(‘<event log name.csv>’.’<eventDate>’),
                TARGET(‘<event log name.csv>’.’<eventDate>’)), 0.25)’
                        Formula Title = ‘Q1’
                        Pre-defined formats = ‘Decimal Number #.##’
                        Units = ‘days’, click done

3.  To create the median:
        Click on New Connection KPI
        In the EDIT FORMULA box type:
                MEDIAN(1.0 *
                DATEDIFF(dd,
                SOURCE(‘<event log name.csv>’.’<eventDate>’),
                TARGET(‘<event log name.csv>’.’<eventDate>’)))
                        Formula Title = ‘Q2’
                        Pre-defined formats = ‘Decimal Number #.##’
                        Units = ‘days’, click done

4.  To create the third quartile:
        Click on New Connection KPI
        In the EDIT FORMULA box type:
                QUANTILE(1.0 *
                DATEDIFF(dd,
                SOURCE(‘<event log name.csv>’.’<eventDate>’),
                TARGET(‘<event log name.csv>’.’<eventDate>’)), 0.75)
                        Formula Title = ‘Q3’
                        Pre-defined formats = ‘Decimal Number #.##’
                        Units = ‘days’, click done

5.  Click the Save button and click Done
```

Figure 4.4 presents a process model example after selecting Custom KPI 2.

Figure 4.4 Custom KPI 2 displaying time intervals between activities


Process Start
208
referral open
208
34.88 Days | 55.65 Days | 73.05 Days
New O/P appt attended
208
21.07 Days | 28.00 Days | 35.08 Days
Follow-up O/P appt attended
205
Process End
208

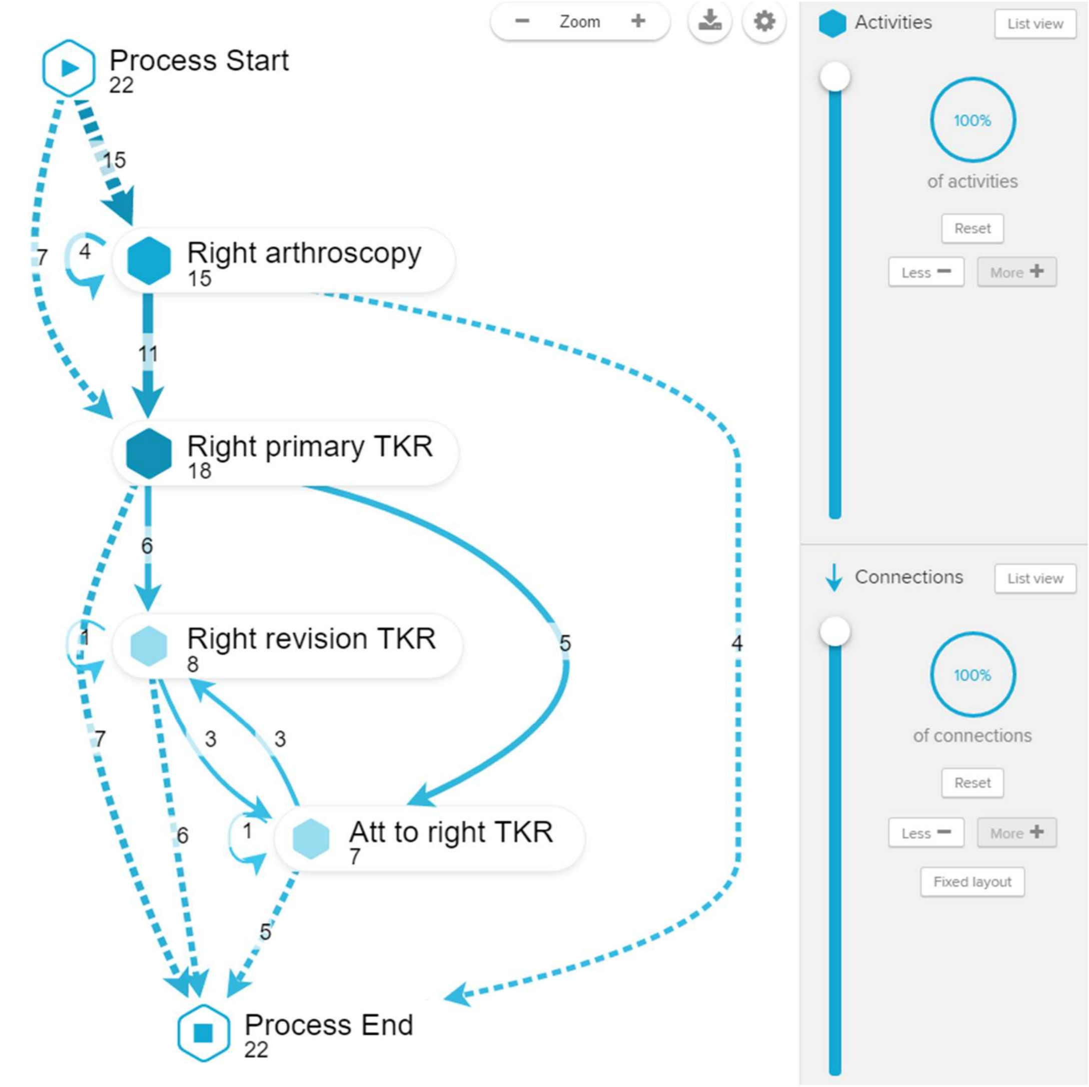
4.1.3.2 Process discovery

In this research programme process discovery is performed using the two process mining tools Celonis and DISCO. The decision of which process mining tool to use should be based on the capabilities of both the algorithm and the features within the tool to handle the imported data, as well as the appropriateness of the output to be used in the following stage of the study. The purpose of process discovery is to reproduce the event log data as closely as possible in model form. It is not to create generalisable models for use with data not yet observed by the model. Celonis and DISCO are extremely similar, in terms of the way that they work and the features they offer. Both tools are based on the Fuzzy Miner algorithm (Section 3.4.5) which is ideal for discovering process models of high quality from complex, messy datasets.

Ease of model reproduction should always be a consideration when undertaking research projects, to allow for future research using different datasets.

The first stage of process discovery when using Celonis and DISCO is to import the event log, then assign a case ID, an event label and a timestamp to the imported data. Process models are then generated from the data, as seen in Figures 4.5 and 4.6.

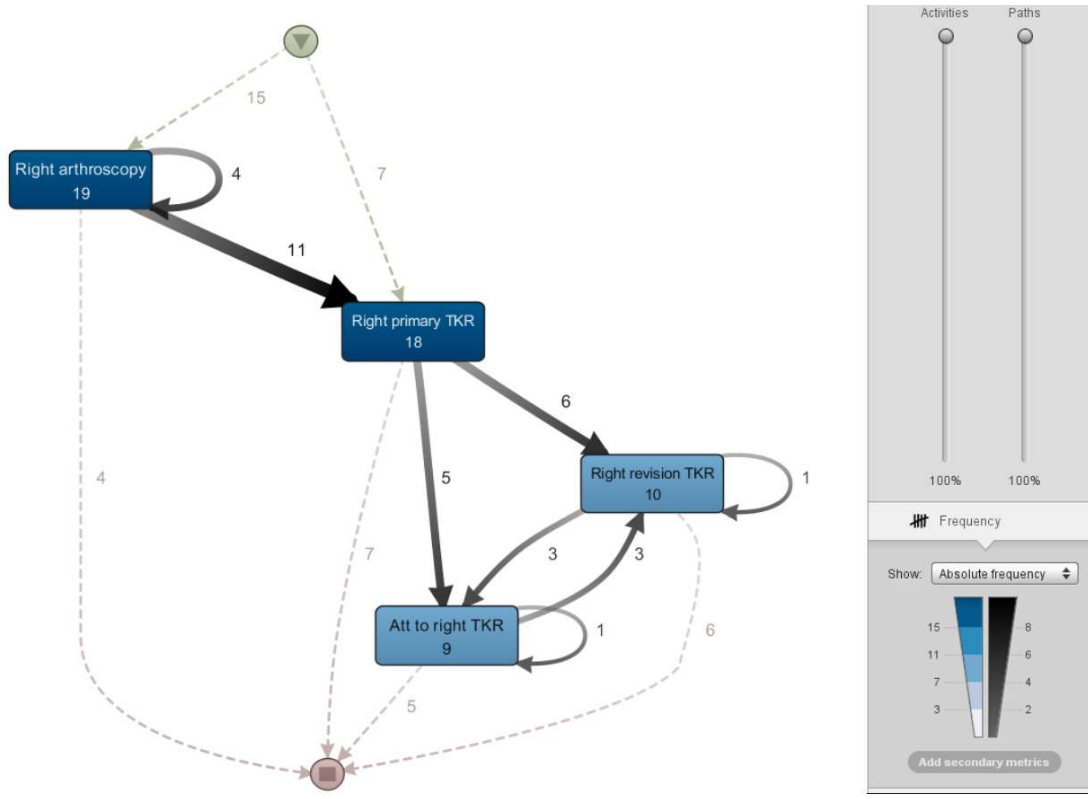
Figure 4.5 Celonis process model example



Events, also known as ‘activities’ in Celonis are represented as blue hexagonal nodes. The activity name is shown beside the activity, with either the case or activity (absolute) frequency shown below. The darker the node, represents the higher the number of cases. Connections between activities are represented by arcs, where the thickness of the arc represents either the frequency or the duration for cases travelling between activities. This information is also shown as a number against the path. Performance statistics may be displayed along the arcs indicating either the median, mean or trimmed mean duration between activities. Dotted lines indicate that the activity is either a starting or ending activity for a case. The number of activities and connections displayed can be adjusted using the sliders to the right. Using Celonis it is simple to filter cases by their activities, process flows or attributes.

The DISCO process model presented in Figure 4.6 is discovered from the same event log as the Celonis model above.

Figure 4.6 DISCO process model example



Show: Absolute frequency

The two discovered models are extremely similar. In DISCO, events, known as ‘activities’ are represented as blue rectangular nodes. The activity name is shown inside the activity, with either the absolute frequency, case frequency or maximum repetitions shown below. Alternatively performance statistics may be displayed below the activity name (start and end times must exist in the event log) indicating either the total, median, mean, maximum and minimum duration. The darker the node, the higher the number of cases. Connections between activities are shown by arcs, known as ‘paths’ in DISCO, where the thickness of the path represents either the frequency or the duration for cases travelling between activities. This is also shown as a number against the path. Faint dotted lines, indicate that the activity can be a starting or final activity for a case. The number of activities and paths displayed can be adjusted using the sliders to the right. As with Celonis, it is simple to filter cases by their activities, process flows or attributes using DISCO.

4.1.3.3 Conformance checking

As stated in Section 3.4.2, conformance checking is performed to reveal the differences between prescribed behaviour, as defined in the reference model, and the recorded executions captured in the event log. In this thesis, conformance checking was performed on the SAIL data in Chapter 7. The ProM framework was used as it provides flexibility in terms of software options, provided via different plugins. The aim of conformance checking was to: 1) identify data cleansing issues, ensuring the data validity prior to analysis; and 2) identify errors in an expert-defined reference model. The inputs needed for conformance checking are a Petri net model and an event log. The ProM plugins used to create a Petri net model in this study are described below.

Mine Petri net with Inductive Miner plugin

Although it is well documented that the Inductive Miner algorithm is capable of discovering sound models with a high replay fitness [351], all Petri net discovery plugins within ProM were tested. The ‘Mine Petri net with Inductive Miner’ plugin [351] was chosen as it generated the model that most closely represented the data. To use the plugin the following steps were taken:

1. Load the synthetic data into ProM in CSV format
2. Convert the data into a log file using ‘Convert CSV to XES’ plugin

3. Select the log file and apply the *'Mine Petri net with Inductive Miner'* plugin with the following settings: Variant = 'Inductive Miner – infrequent (IMf)'; Event Classifier = 'Event Name'; Noise threshold = 0 (for perfect log fitness).

4. The discovered Petri net was visualised using 'Visualize Petri net (PetriNet)' visualisation.

Mine Process Tree with Inductive Miner plugin

The Inductive Miner algorithm is used with this plugin [351] to discover and edit a process tree. Steps a to c above were repeated using the 'Mine Process Tree with Inductive Miner' plugin and the process tree was visualised using the 'Graphviz Process Tree (Inductive Visual Miner)' visualisation.

The Petri net model and event log were then used for conformance checking using the Multi-Perspective Process Explorer (MPE) [252] plugin.

Multi-Perspective Process Explorer (MPE) plugin

The MPE provides five main views on the event log data. For each view apart from Model mode, the percentage and number of violations is displayed. Violations are separated into two types, wrong and missing events. A 'Missing Event' indicates there has been a move on the model only, for example, a revision to surgery, with no prior surgery. A 'Wrong Event' indicates a move in the event log only, an example of this would be where a patient has two right-sided primary surgeries. Detailed analysis of the violations can then be performed using the trace view. For all views apart from Model mode, the event log data can be filtered. The filtering options include the logical operators 'AND', 'OR' and 'NOT'.

- **Model mode:**
  Displays the overall statistics including the number of patients in the event log and the number of events.

- **Fitness mode:**
  Displays the average fitness between the event log data and the model.

- **Data discovery mode:**
  Displays the frequency of cases over the different routes taken through the model.

- **Performance mode:**
  Allows for the data to be explored by displaying the two measures against each arc on the model. These measures are configurable and consist of: percentage (trace); average time; minimum time; maximum time; median time; first quartile (Q1) time; and third quartile (Q3) time.

- **Precision mode:**
  Displays the average activity precision along with the number of moves observed and the number of moves possible.

4.1.3.4 Enhancement

In Chapter 7, enhancement activities were performed in order to repair the meta data of existing process model using executions from the event log. During the conformance checking, three new low-level surgery codes were identified and added to the meta-data of the reference model. This meta-data was used when extracting events from the SAIL data.

4.1.4 Stage 4: Evaluation

The objective of the evaluation stage was to ensure that all study-specific research questions had been answered using the results from each study and that the results were meaningful to the domain experts. In addition, interesting and unusual observations, limitations and potential further or future work were identified. Evaluation exercises were performed throughout the various stages to assess the approach and to gather input from the domain experts. At the end of each study, formal meetings were held with domain experts to validate and verify the results.

Evaluation of the results in this research programme have been conducted in three ways: 1) by comparison with those from similar peer reviewed studies; 2) by measurement of process model quality, using the standard metrics described in Section 3.4.1; and 3) by clinical domain experts. Clinical evaluation of the study results in Chapter 5 was performed by Dr Klaus Witte, a consultant cardiologist, for Chapter 6 by Mrs Helen Wilson, a Clinical Service Manager and Enhanced Role Physiotherapist and for Chapter 7, by Professor Hemant Pandit, an orthopaedic surgeon, Professor Philip Conaghan, a rheumatologist from Leeds Teaching Hospitals NHS Trust (LTHT) and Dr Mar Pujades Rodriguez, a medical epidemiologist from Union Chimique Belge Biopharma. Technical verification was carried out for all studies through discussions with project team member Mr Owen Johnson. Epidemiological evaluation was performed by Dr Rodriguez. Finally, evaluation by systems experts was carried out by discussions with Mr Simon Bramwell, a business analyst at ADI for Chapter 6, and various members of the SAIL analysis team for Chapter 7. In addition to the evaluation of results, clinical and technical verification was carried out during each stage through discussions with the project team members.

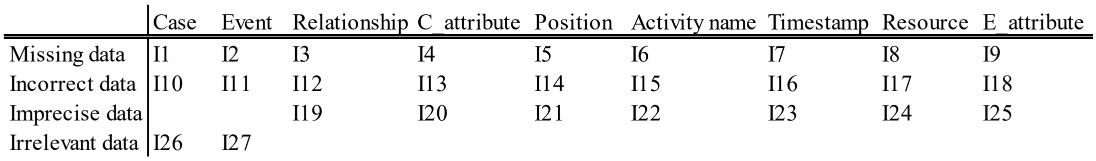
4.1.4.1 The role and methods of evaluation

Different verification and validation techniques appropriate for consideration in this thesis have been discussed in Section 3.6. The role of evaluation during research projects is critical to ensure: 1) the research questions asked are relevant; 2) the methods and techniques applied are capable of producing high quality artefacts that are correct, understandable and usable; and 3) the results, findings and conclusions generated from the artefacts are correct, accurate, useful and realistic. The following two sections describe two evaluation techniques that were applied during this work.

4.1.4.1.1 Data quality evaluation technique

In Chapter 6, the quality of the final data extract was evaluated using the Mans Data Quality Matrix, introduced in Section 3.6.1. Table 4.2 presents the matrix with the four data quality issue types: Missing; Incorrect; Imprecise; and Irrelevant. These issues may be applicable to the following entities in an event log: Case; Event; Relationship; Case attribute; Position; Activity name; Timestamp; Resource; and Event attribute.

Table 4. 2 Mans Data Quality Matrix [287]



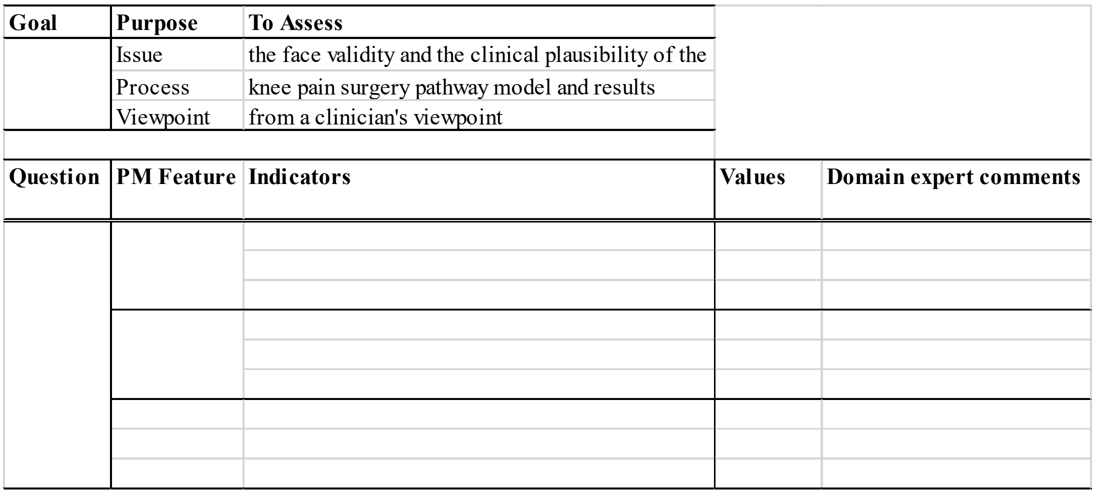
	Case	Event	Relationship	C_attribute	Position	Activity name	Timestamp	Resource	E_attribute
Missing data	I1	I2	I3	I4	I5	I6	I7	I8	I9
Incorrect data	I10	I11	I12	I13	I14	I15	I16	I17	I18
Imprecise data			I19	I20	I21	I22	I23	I24	I25
Irrelevant data	I26	I27							

When assessing data quality using the matrix, issues with a value of ‘N’ indicate no issue, ‘L’ indicate a low infrequency and ‘H’ indicate a high frequency within the data. Cells left blank, indicate that the issue does not apply.

4.1.4.1.2 Interview evaluation technique

The GQFI Table introduced in Section 3.6.3.4 was revised and used in Chapter 7 to structure an interview with clinical domain experts. The table was extended for use in this study by adding two further columns, ‘Values’ and ‘Domain expert comments’. This extension allowed for the results and values of each KPI to be presented, and for the validation carried out against those values to be evidenced. Additionally, rather than ‘Goal’ referring to the goal of the project, it refers to the goal of the interview. The ‘To Understand’ column was replaced with ‘To Assess’, in alignment with its purpose. The amended table is presented below.

Table 4. 3 GQFI Table



Goal	**Purpose**	**To Assess**		
	Issue	the face validity and the clinical plausibility of the		
	Process	knee pain surgery pathway model and results		
	Viewpoint	from a clinician's viewpoint		
Question	**PM Feature**	**Indicators**	**Values**	**Domain expert comments**

This table provides a clear, summarised view of the results for discussion during the interview. All evidence collected during the technical evaluation and after comparison with published scientific literature is entered prior to the interview.

The goal of the interview is structured using the headings 'Issue', 'Process' and 'Viewpoint' to provide all participants with a clear understanding of the purpose of the interview. Research questions and process mining features are used to answer the questions and KPIs are recorded in the main body of the table. An example is:

- **Question:** '*Can useful healthcare statistics be generated from the SAIL data for patients with knee pain using process mining techniques?*'
- **PM Feature**: '*Process variant analysis*'
- **Indicator**: '*Percentage of patients having only right primary TKR surgery*'
- **Value**: '*21%*'

This approach presents the results to the domain experts in a clear and concise way and provides a framework for structuring the interview, ensuring that all necessary information is captured. Comments from the experts are entered into the final column. It is important that questions to help evaluate the clinical plausibility of the results are included but information must also be gathered on the relevance and usefulness of the results along with the understandability and usability of any artefacts.

4.1.5 Stage 5: Process improvement

The process improvement stage is more commonly carried out for industry-based projects, as the emphasis is often on improving the performance of business processes for cost saving purposes. Process improvement activities were not carried out for the studies presented in chapters 5 and 7. However, process improvements resulted from the second study and these are reported in Chapter 6.

4.2 The research datasets

Three separate studies were undertaken as part of this research programme. The first study in Chapter 5 uses MIMIC-III data to establish whether process mining techniques can be effectively used to create disease trajectory models. Here we attempt to replicate the results of Jensen et al. who used statistical techniques to produce disease trajectory models using registry data for the entire Danish population (Section 5.2). Results for the Danish study showed gout to be a central diagnosis within a cluster of patients diagnosed with CVD.

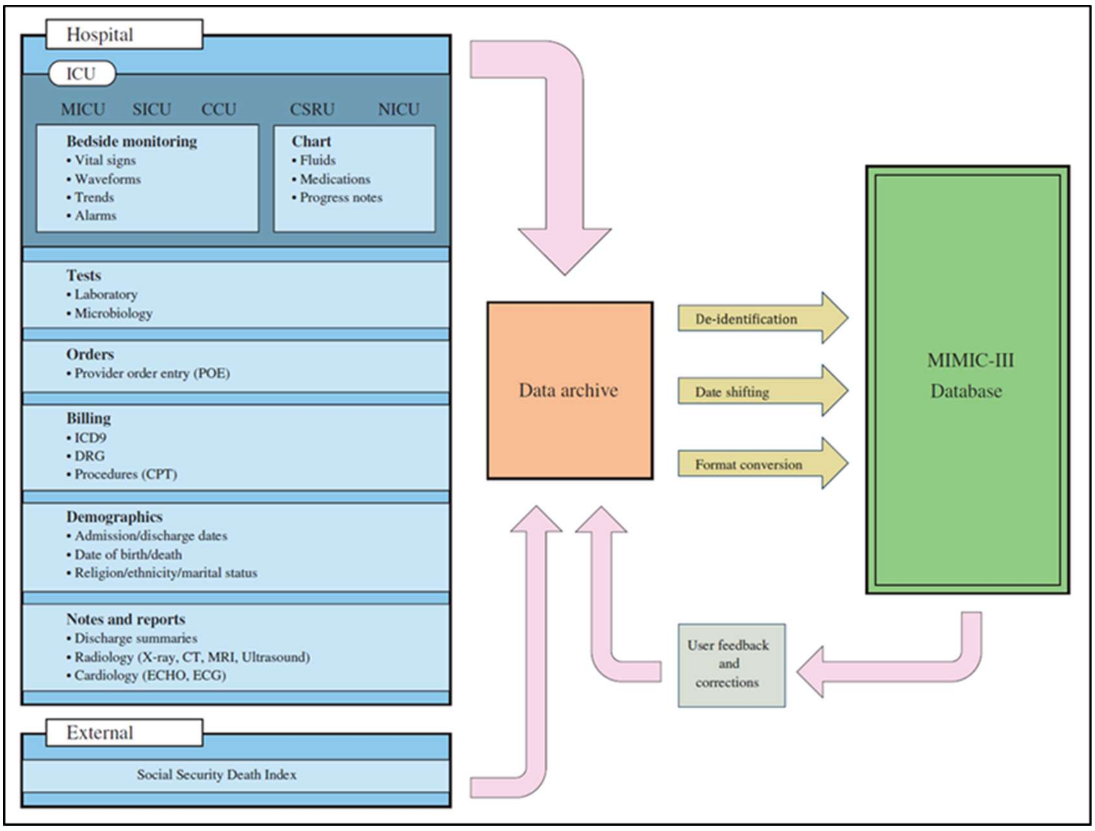
The second study in Chapter 6 uses process mining techniques to discover patterns of care for groups of patients with different patient-reported outcomes. The outcome measures were based on the analysis of patient-reported data, collected using the MyPathway application for patients attending the PhysioWorks service within Sheffield Teaching Hospitals NHS Foundation Trust.

The final study in Chapter 7, again uses process mining techniques to help create an expert-defined, interactive reference model for knee pain surgery using data from the SAIL databank. This reference model is then used to check how the actual SAIL data conforms to the reference model. Finally, the SAIL data was used to generate episode statistics for knee pain surgery. The following three sub-sections expand on the information provided in Section 1.4.

4.2.1 The Medical Information Mart for Intensive Care III dataset

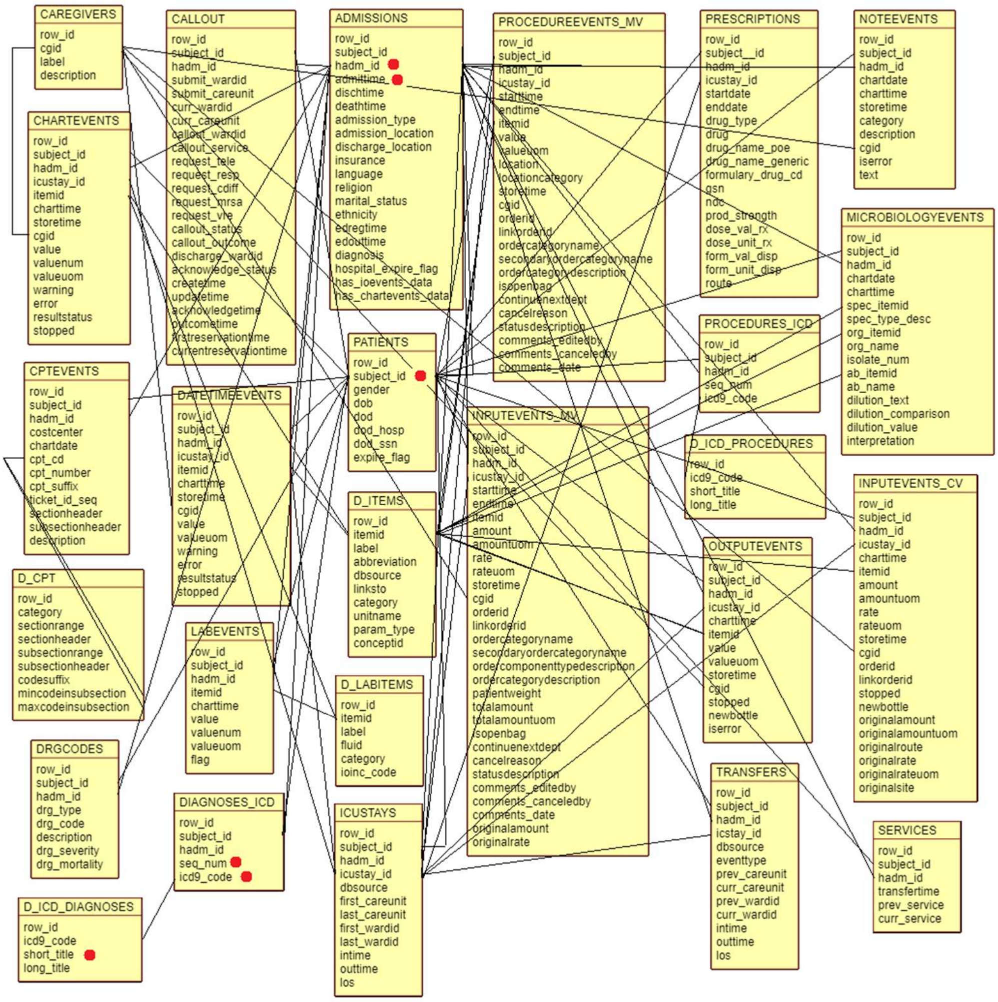
Hospital data from the MIMIC-III database version 1.4 [17] was used for this study. Specifically, data for patients diagnosed with CVD, where the validity is classified as high and sufficient for use in research [352]. MIMIC-III is a large, freely available database consisting of 46,520 distinct de-identified patient records and 58,976 hospital admissions. The data architecture diagram in Figure 4.7 gives a high-level view of how the clinical patient information is stored including demographics, charted observations with vital sign measurements taken at the bedside, procedures, in and out patient laboratory test results, medications, caregiver notes including discharge summaries, imaging reports, and mortality.

Figure 4.7 Data architecture for the MIMIC-III database [33]



Electronic health records are made up of 38,645 adult (age 16 and above) records (recorded between 2001 and 2012) and 7,875 neonate records (recorded between 2001 and 2008). MIMIC-III is a relational database consisting of 26 tables, as seen in Figure 4.8, where all data uses the International Classification of Diseases (ICD-9) [353] format. To access to the data, it was necessary to reconstruct the database onto a local machine. Twenty-six comma-separated values (CSV) files, totalling 6.2 GB in size, were provided via the Physionet website [354], along with build and import scripts. These scripts and files were used to create a local copy using the object-relational database management system PostgreSQL 9.6. In Figure 4.8, data items used in the final data extract in Chapter 5 are highlighted with a red dot.

Figure 4.8 MIMIC-III data model



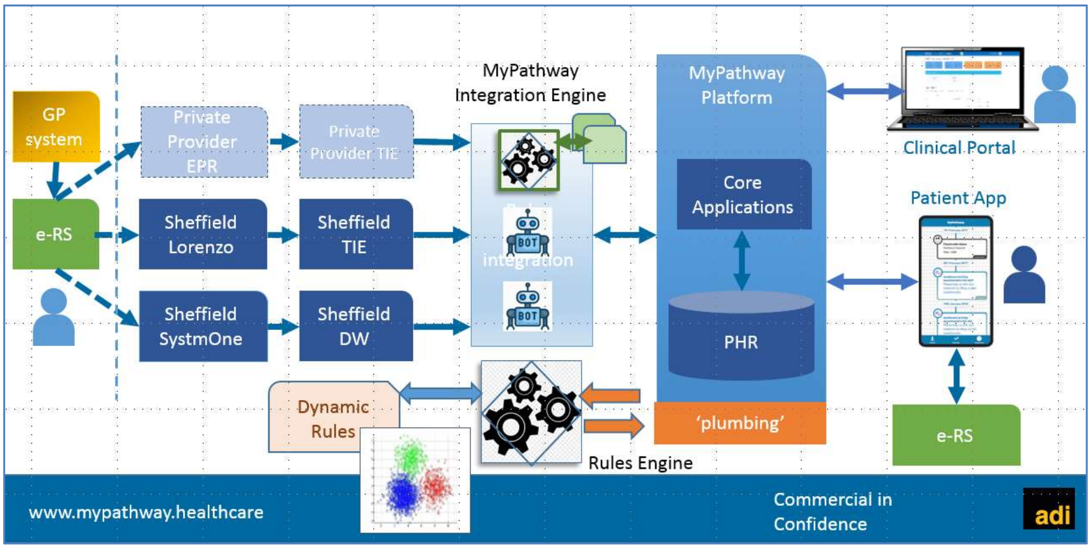
Data is extracted from various sources which include two different critical care information systems (CCIS) within the hospital, CareView and MetaVision. The two CCISs store and display clinical data at the bedside for intensive care unit patients. However, CareView stores data for patients admitted between 2001 and 2008 whereas MetaVision stores data for patients admitted from 2008 onwards. A consequence of having these two systems, is that the information is stored in different formats. Therefore, to extract data for the entire duration of the database different identifiers are required. Where possible, similar data from both of these systems is combined into one MIMIC-III table, but where this was not possible separate tables with either the suffix *cv* or *mv* were created. The Social Security Administration Death Master File was used to extract out-of-hospital mortality dates. Finally, data was extracted from the hospital and laboratory EHR databases for information such as patient demographics, in-hospital mortality, laboratory test results, discharge summaries and reports of imaging studies. Records are linked using either *subject_id* which refers to the patient, *hadm_id* which refers to a patient admission or *icustay_id* which refers to an intensive care unit admission. A patient's stay can be tracked using the following tables: *ADMISSIONS, PATIENTS, ICUSTAYS, SERVICES* and *TRANSFERS*.

All dates, including dates of birth (DOB) have been randomly shifted into the future to assist with patient confidentiality, though all dates for the same patient are internally consistent. If the patient is older than 89 years their date of birth is set to 300 for their first admission. Times are stored to the minute and dates to the day, therefore measurements stored with the suffix *date* will have a time of '00:00:00' this does not indicate that the measurement was taken at midnight, rather the time has not been recorded. *charttime* and *storetime* are used to record patient observations. *charttime* is when the observation was actually taken from the patient and will be recorded usually to the previous hour, e.g. a measurement taken at 14:23 will be recorded as 14:00. *storetime* is when the measurement has been validated and stored in the database, this is recorded down to the number of minutes and is usually within a number of hours of the measurement being taken. *deathtime* is the date and time of the patient's death, only if they died in hospital, else this is null.

4.2.2 The MyPathway dataset

MyPathway is a cloud-based application accessible to patients and health professionals via a web browser using a computer or mobile device. The patient health record (PHR) and all related MyPathway information is stored in a Mongo NoSQL database [355] (Section 3.1.2) using a JSON-like document based structure [356]. This modern type of database structure is extremely fast and flexible and can quickly be amended to meet changing business requirements. However, drawbacks do exist, especially when systems have been designed without reporting capabilities in mind, these drawbacks are discussed in Chapter 6. The implementation of the MyPathway system for STHT is shown below in Figure 4.9.

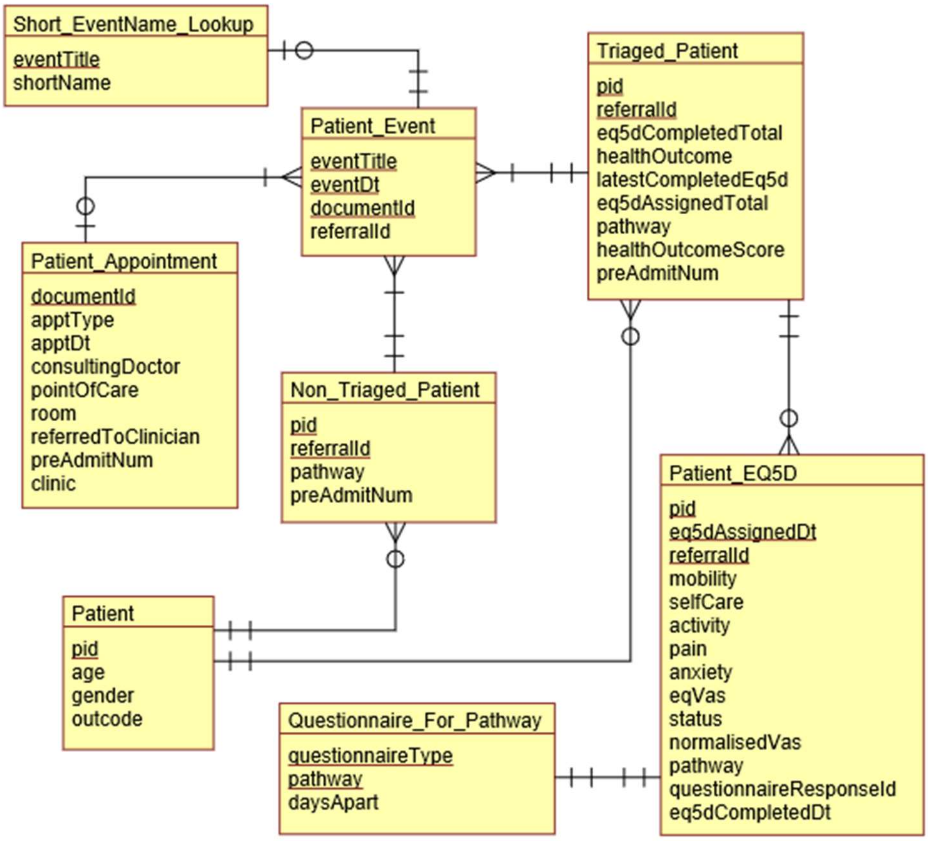
Figure 4.9 MyPathway system architecture diagram as implemented by STHT [357]



Here, patient data is extracted from the Sheffield primary and secondary care information systems Lorenzo and SystmOne and from the NHS e-Referral Service (e-RS) for use in the MyPathway application. However, a limitation to this study is that the patient diagnosis codes were not included. The MyPathway integration engine uses pre-defined rules to automatically trigger events for patients within the system. These rules may be time, patient or condition dependant. In addition to these automatically triggered events, clinicians and patients may initiate and respond to activity within the system by using the clinical portal and patient app.

As Mongo databases use a JSON-like document based structure with no concept of relational linkage, a logical data model was created by the author for data extraction purposes and is shown in Figure 4.10.

Figure 4.10 Required logical data model for data extracted from the Mongo database



The dataset relates to patients referred to STHT by their GP between 15/05/2017 and 12/08/2019 and contains 119,266 de-identified patient records. The data model shows patient information which may relate to either a triaged or non-triaged patient referral. Triaged patient referrals contain EQ-5D questionnaire information, which may include patient response data. Patients also have associated events including appointment events. Along with basic patient appointment information such as date and time, other information is recorded to enable data cleansing and linkage activities to be performed during the data transformation stage. Access to the data was provided by ADI's technical staff in the form of 156 csv files, each containing 23 data items which related to the attributes shown in Figure 4.10.

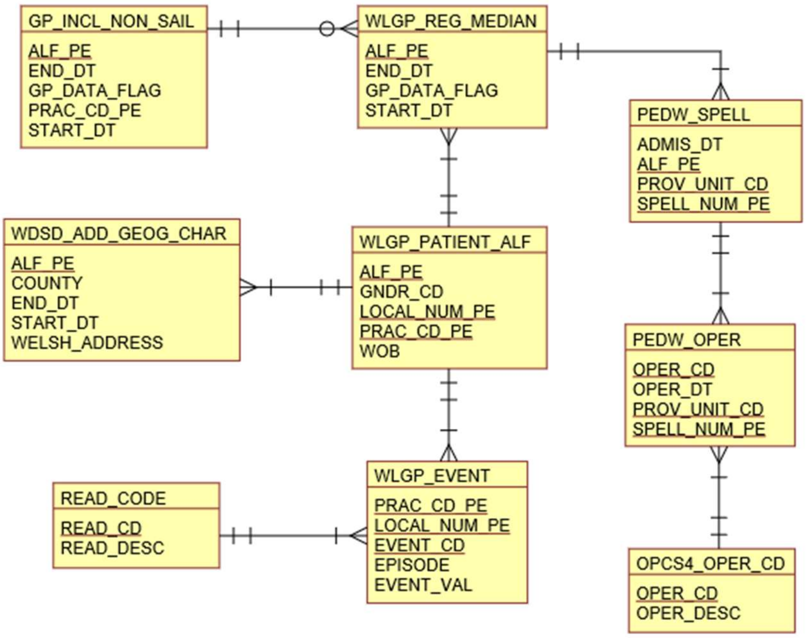
4.2.3 The SAIL dataset

The SAIL databank is a large database containing over 15 billion anonymised records, relating to between 4 and 5 million people [30]. Due to strong data security requirements, the data is hosted on a cloud platform located in Wales and must be accessed via a virtual private network (VPN). A VM Horizon client with a mobile authentication devise is used to access a dedicated gateway on a virtual machine. The gateway provides access to a dedicated desktop where the data stored in a DB2 database can be accessed through the Eclipse IDE. General software was provided, though research specific tools such as the ProM framework, Celonis and DISCO needed to be installed by the researcher.

The three SAIL datasets introduced in Section 1.4.2 (WLGP, WDS and PEDW) are further described in this section. Both the WLGP and the WDS datasets are provided using the Audit+ software, which has been provided free of charge to all GP practices in Wales. Two anonymised files are automatically extracted and securely transferred into SAIL containing patient demographics (WDS) and clinical event details (WLGP). Hospital data (PEDW) is collected via the Admitted Patient Care Data Set (APC Ds) [358] from each NHS health care provider in Wales. This is provided in a fixed format file via a secure upload mechanism on the NHS Wales Data Switching Service (NWDSS).

Figure 4.11 shows the entity relationship diagram containing the SAIL views used in this study. All identifiers have been encrypted for anonymization purposes. The *ALF_PE* identifier is used to link patients across all datasets. Within the WLGP dataset the linking fields *PRAC_CD_PE* (unique surgery code) and *LOCAL_NUM_PE* (unique patient code) are used in place of the *ALF_PE.* The PEDW dataset uses *PROV_UNIT_CD* (unique hospital code) and *SPELL_NUM_PE* for linkage. GP diagnoses are recorded using Readcode version 2 and descriptions for these diagnoses are held in the READ_CODE lookup table. Up to 12 operations can be recorded for an episode of care using the OPCS-4 codes. Descriptions for theses operation codes are held in the OPCS4_OPER_CD lookup table.

Figure 4.11 Entity relationship diagram for SAIL views used



All events in SAIL are stored in date format only ‘*yyyy-MM-dd’*. As patient data is anonymised, a patient’s date of birth is represented as *week of birth* (WOB) and is calculated using the date of the Monday that occurs prior to the date of birth.

Validity checking was carried out by querying and joining the tables to ensure the results were plausible in their context. Analysis of the PATIENT_ALF table was used to validate WOB. Results found that 0.01% (n=228) of patients were over the age of 110 (with no recorded death date) or under the age of 0 years at the end of the study period. These patients were not included in any of the analysis. The recording of operations in the WLGP dataset is inaccurate. The number of TKR events in the GP_EVENT table between 01/01/2000 and 30/09/2017 (n=18,886) was compared against the number in the OPER table using the same dates (n=67,295). Due to this discrepancy, all operation data was selected from the PEDW dataset. After joining to the CLEAN_ADD_GEOG_CHAR table it was found that 21,575 patients were not resident and received care whilst travelling through Wales. These patients were excluded from the analysis.

Chapter 5
Process mining MIMIC-III data for disease trajectories

5.1 Introduction

Due to an increased prevalence and burden of long term conditions, new methods for managing and delivering health care are urgently needed. EHR data has the potential to reveal insights into the biomechanics of disease progression or trajectories over time. A better understanding by medical practitioners of disease progression and information gained from the study of care pathways could enhance patient health outcomes. Care pathways can be used to help understand disease development and to provide early diagnosis within clusters of diseases, helping to mitigate the risk of adverse health outcomes. In combination with other methods, outputs from disease trajectory models could be used by healthcare professionals to target patients for appropriate, early primary prophylaxis. This may help to prevent or delay disease development, for patients who require more intensive or aggressive therapy, or for those who may benefit from a more expensive, but safer medication.

A method that may be used alongside disease trajectory modelling is prediction modelling. Disease trajectory models use historic data to describe what has happened in the past. They are diagrams based on fact and real life. Disease trajectory models may be used by clinicians to test hypotheses or as the basis for deciding on prediction. Prediction is a separate science where models are created using statistical techniques to estimate the probability of an outcome [359] based on a set of inputs. The two types of model are used for different purposes, though can be used in combination in order to provide a more in-depth understanding based on both reality and probability.

The work in this chapter has pioneered the use of process mining techniques for disease trajectory modelling [329], [360]–[362]. Inspiration for this study was taken from work published by Jensen et al. [38] in 2014, where statistical methods were used to create a hand-drawn disease trajectory model for a CVD cluster of diseases. Results from this study showed gout to be a central disease within this cluster. Later in 2019, Chen-Xu et al. commented on how a rise in prevalence of comorbidities and obesity is thought to be a main contributor to the increase in gout cases worldwide and how collaborative efforts are required to help improve widespread sub-optimal management [198]. In this chapter Jensen's rules are followed in order to create a set of iteratively refined CVD disease trajectory models by applying a process mining approach using the MIMIC-III dataset. Improvements to the Jensen method are proposed and described by illustrating how a process mining approach can help to advance the study of disease trajectories from data stored in EHRs.

5.2 Stage 1: Planning

During the planning stage the Jensen [38] article was found and a decision was made to replicate the CVD trajectory rules to allow for the model to be reproduced using process mining techniques. The type of data available from the MIMIC-III database was reviewed by visiting the official website [363] and by reading the recommended literature [17], [18] to obtain an understanding of the data (see Section 4.2.1). In addition to the core project team, defined in Section 4.1.1, an external programming resource was used to assist with the writing of the data transformation code. Approval to use the data was granted [17] by PhysioNet [32] after successful completion of the online Collaborative Institutional Training Initiative (CITI) 'Data or Specimens Only Research' course.

This study aims to establish whether process mining techniques can be effectively used to create disease trajectory models using the MIMIC-III data. Five study-specific research questions were composed towards this aim and as an implementation of the primary research questions in Chapter 1, these consisted of the following:

1. Is it possible to create an event log, using data for patients diagnosed with cardiovascular diseases, from the MIMIC-III data?

2. Can process mining techniques be used to create disease trajectory models?

3. Can the cardiovascular disease trajectory results reported by Jensen be reproduced using the MIMIC-III data and a process mining approach?

4. What differences exist between Jensen’s trajectory model and the model created using the MIMIC-III data?

5. What are the strengths and weaknesses of using a process mining approach to construct disease trajectory models?

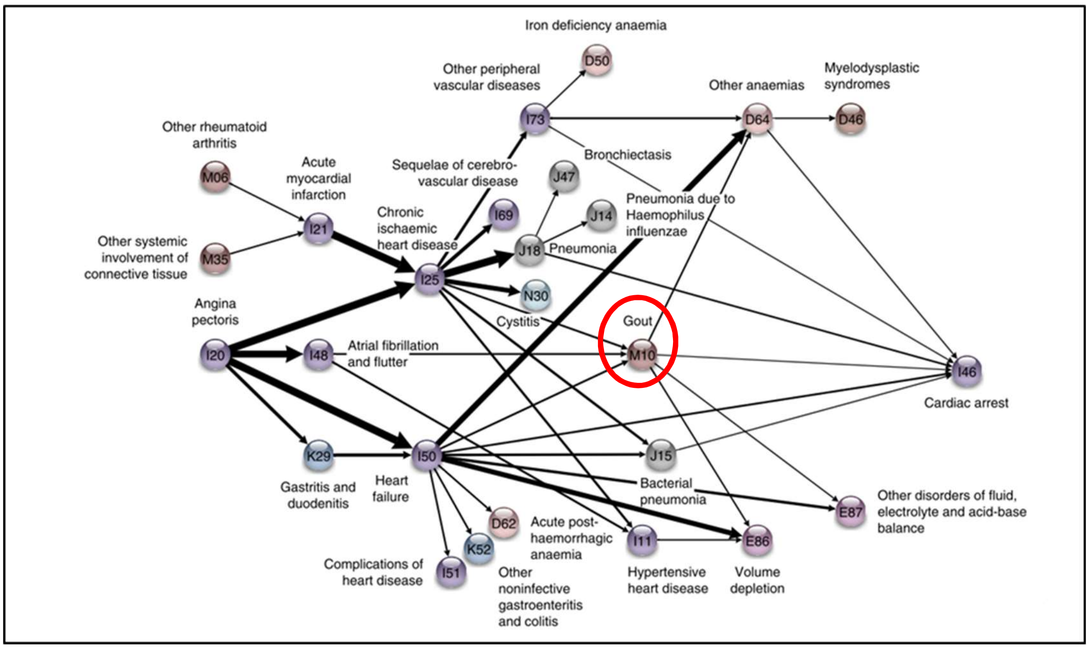
5.3 Stage 2: Extract, Transform and Load

Stage 2 began with the gathering of detailed methodological knowledge by carefully extracting and interpreting Jensen’s rules from the published works. Communication with the author [75] was necessary to clarify details not present in [38]. Before extraction of the MIMIC-III data could take place it was necessary to reconstruct the dataset onto a local computer. The data was extracted using SQL, before it was transformed and loaded into the process mining tool DISCO.

5.3.1 Overview of the Jensen method

The CVD trajectory model showing gout as a central diagnosis is presented in Figure 5.1.

Figure 5.1 Cardiovascular disease trajectory model by Jensen et al. [38]


Iron deficiency anaemia
D50
Other peripheral vascular diseases
I73
Other anaemias
D64
Myelodysplastic syndromes
D46
Other rheumatoid arthritis
M06
Bronchiectasis
Acute myocardial infarction
Sequelae of cerebro-vascular disease
J47
I69
Pneumonia due to Haemophilus influenzae
J14
Chronic ischaemic heart disease
I21
Other systemic involvement of connective tissue
M35
J18
Pneumonia
I25
N30
Cystitis
Angina pectoris
Gout
I20
I48
Atrial fibrillation and flutter
M10
I46
Cardiac arrest
K29
I50
J15
Gastritis and duodenitis
Heart failure
Bacterial pneumonia
E87
Other disorders of fluid, electrolyte and acid-base balance
D62
Acute post-haemorrhagic anaemia
K52
I11
E86
Complications of heart disease
I51
Other noninfective gastroenteritis and colitis
Hypertensive heart disease
Volume depletion

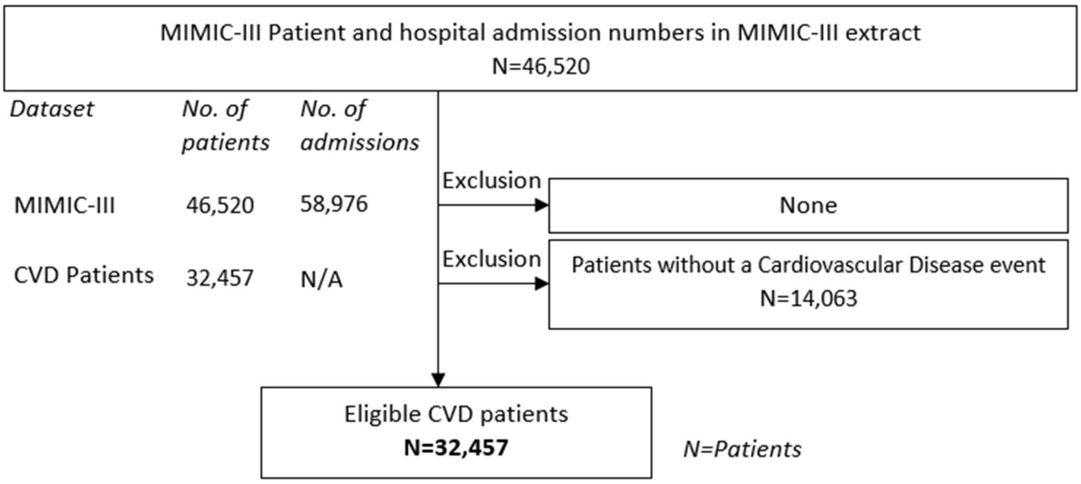
The model consists of coloured nodes representing diseases and directed arcs representing common trajectories between diseases. Diseases are represented using their associated ICD-10 code and the colour of the node indicates the ICD-10 chapter. The thickness of the arc represent the relative number of patients. Jensen used secondary care data from the Danish National Patient Registry (NPR) between 1996 and 2010 to create the trajectory model. This data originated from approximately 89 public and private hospitals, covering the entire population of Denmark, approximately 6.2 million.

The data selection rules defined in [38] were used as a basis for creating the disease trajectory models in this study. All data was coded using the ICD-10 [353] hierarchical structure and rounded down from level five to level three. Any codes within the following chapters were excluded: pregnancy related; symptoms, signs, and ill-defined conditions; injury and poisoning; and supplementary classification of factors influencing health. Data was stratified by hospital encounter type consisting of inpatient admissions, outpatient and emergency room visits. Further clustering was carried out to form groups with similar types of diseases, such as the CVD cluster which is used in this study. Within the CVD cluster, statistically significant temporal directional pairs of diseases (bi-grams) were created. There were three main stages to Jensen's experiments: 1) Estimation of association of all pairs of diseases with relative risk derived from a sampling process; 2) Testing of directionality between pairs using a binomial test; 3) Creation of trajectories through the combination of directional pairs, then allocation of patient counts to each trajectory. A restriction was placed on this step whereby for a trajectory to be valid, a minimum of four diagnoses in more than one hospital episode needed to exist. Jensen defined a disease trajectory as set of sequential disease associations. For cases where multiple diagnoses were assigned in the same discharge, the order in which were recorded was considered to be correct. Repeating patient diagnoses were not eliminated, as they did not want to make the assumption that the first time a diagnosis appeared in the data, was the first time it appeared in the patient.

5.3.2 Data extraction

The MIMIC-III relational database (Section 4.2.1) was reconstructed onto a local computer, before using the pgAdmin 4 v1.1 development platform [364] to query and extract the data. Data extraction code was written to select a subset of the MIMIC-III data and write it to csv files. The selection criteria for the inclusion was that patients must have at least one CVD code. Within the ICD-9 hierarchy, CVD codes are in Chapter 7 which relates to 'Diseases of the circulatory system' and range between 390 and 459. The number of patients and hospital admissions extracted are presented in the Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) diagram in Figure 5.2.

Figure 5.2 STROBE diagram showing the MIMIC-III patient and hospital admission numbers


MIMIC-III Patient and hospital admission numbers in MIMIC-III extract
N=46,520
Exclusion
None
Exclusion
Patients without a Cardiovascular Disease event
N=14,063
Eligible CVD patients
N=32,457
N=Patients

Dataset	*No. of patients*	*No. of admissions*
MIMIC-III	46,520	58,976
CVD Patients	32,457	N/A

Approximately 70 percent of all patients in the MIMIC-III database had at least one CVD code within their EHR data. As per Jensen method, all transformation rules excluded patients with ICD-9 codes in chapters: 11 and 15, pregnancy related; 16, symptoms, signs and ill-defined conditions; 17, injury and poisoning; and 18 and 19, supplementary classification of factors influencing health. Using the clinical codes within the chapter hierarchy structure provided an efficient way to select the patients. Despite the size and complexity of the MIMIC-III schema, the comprehensive online documentation that has since become available, is a good resource for providing guidance when needing to understand the data items. The following data items were extracted for use during the data transformation stage: patient (subject_id); patient admission (hadm_id); admission time (admittime); diagnosis code (icd9_code); diagnosis description (short_title); and the sequence number (seq_num) associated with the diagnosis code (Section 4.2.1). The total number of patient diagnoses extracted was 422,616, this is due to the fact that each patient can have multiple diagnoses.

5.3.3 Data transformation and load

An incremental approach was taken to replicating Jensen's CVD trajectory model by applying ten different data transformation rules in order to perform two separate experiments. The first experiment was created using the features of Rule 9 and the second using the features of Rule 10. Table 5.1 presents the list of features included in the ten rules.

Table 5.1 Transformation rules

Rule #	Sequence # = 2 for secondary diagnoses	Level 3 ICD-9 codes	Valid TS in load file	Repeat diagnoses excluded	Secondary diagnoses excluded	Secondary diagnoses TS=primary diagnosis TS+1	Time ordered secondary diagnoses (sequence#)	Only patients > 1 admission	Only primary diagnosis in first admission	Only 1 non-repeating diagnosis per admission	Directional ordered pairs
1	X										
2		X									
3		X	X				X				
4		X	X			X					
5		X	X		X			X			
6		X	X	X	X			X			
7		X						X	X	X	
8		X	X	X			X				
9		X	X								X
10		X	X	X					X		X

TS = timestamp

When working with healthcare data it is important to have involvement from the clinical domain experts in order to understand the subtleties and meaning behind the data. For example, cardiovascular diseases are typically chronic in nature and any re-occurring diagnoses in the patient's EHR are likely to be a re-recording of a previous disease rather than a new instance. Due to this, Rule 6 was created to exclude repeating diagnoses for patients.

Rule 1: Sequence numbers in the extract range from 1 to 39. Here, all sequence numbers apart from 1, which represents a primary diagnosis, are changed to 2 indicating that the diagnosis is secondary to the primary reason for admission [365].

Rule 2: The original three, four or five digit ICD-9 diagnosis code is trimmed to three digits. This feature is applied to all rules between 2 and 10.

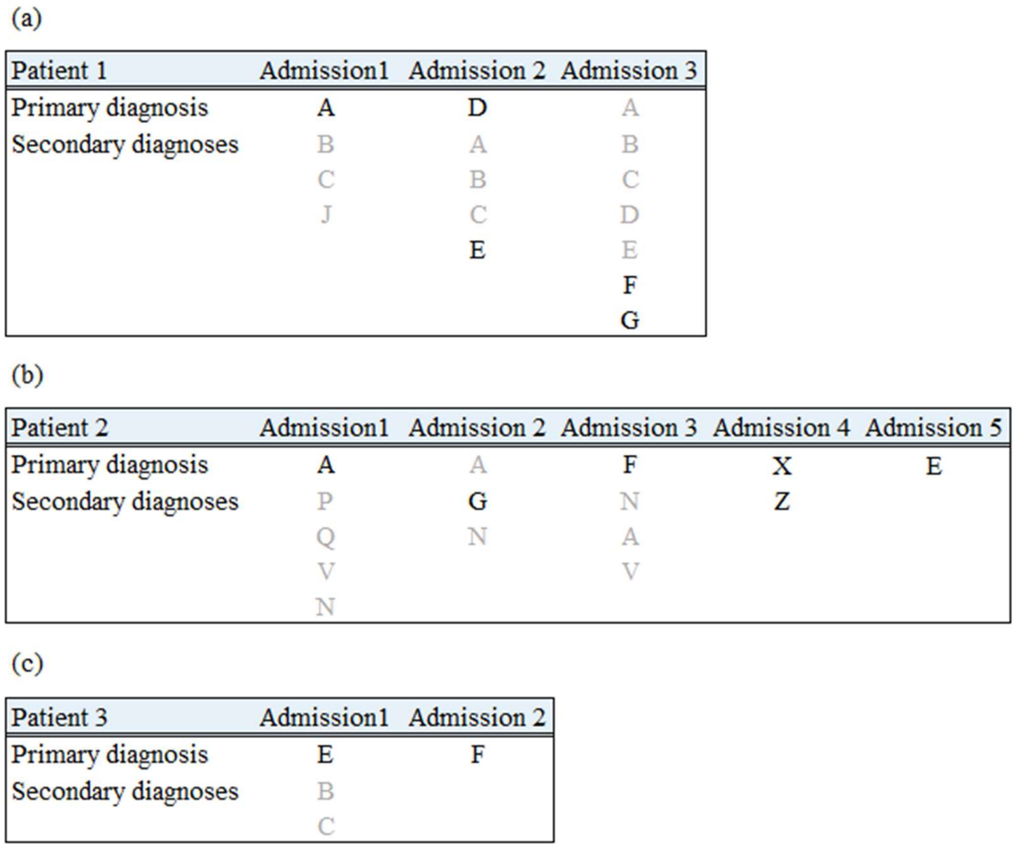
Rule 3: Within the MIMIC-III data, timestamps are not attached to individual diagnoses within hospital admissions, only to the overall admission. The data transformation code allocates the admission timestamp to the primary diagnosis for an admission (sequence number 1). An assumed chronological ordering is then applied to all diagnoses within an admission by allocating a timestamp to secondary diagnoses. For this rule only, this is determined by the sequence number value.

Rule 4: This is similar to rule 3 though, as recommended by Tomp [365], no ordering is allocated amongst the secondary diagnoses. Tomp stated that any sequence number value apart from 1, is used only for billing purposes and does not refer to the order in which a diagnosis was made. Therefore, all secondary diagnoses within an admission are assigned a timestamp one second after the admitted time.

Rule 5: This rule considers only the primary diagnosis within an admission. Therefore, to allow for a disease ‘trajectory’, all patients must have more than one admission within the data.

Rule 6: As with rule 5, only primary diagnoses are considered. In addition, for subsequent patient admissions, repeating diagnoses are excluded. When repeating diagnoses are excluded, only the primary diagnosis is considered for the first hospital admission, as it is unknown whether the secondary diagnoses in the first admission are repeated from a previous manifestation prior to the study window. Figure 5.3 presents three examples of how patient diagnosis data is processed.

Figure 5.3 Processing examples for repeating diagnoses



(a)

Patient 1	Admission1	Admission 2	Admission 3
Primary diagnosis	**A**	**D**	A
Secondary diagnoses	B	A	B
	C	B	C
	J	C	D
		E	E
			F
			G

(b)

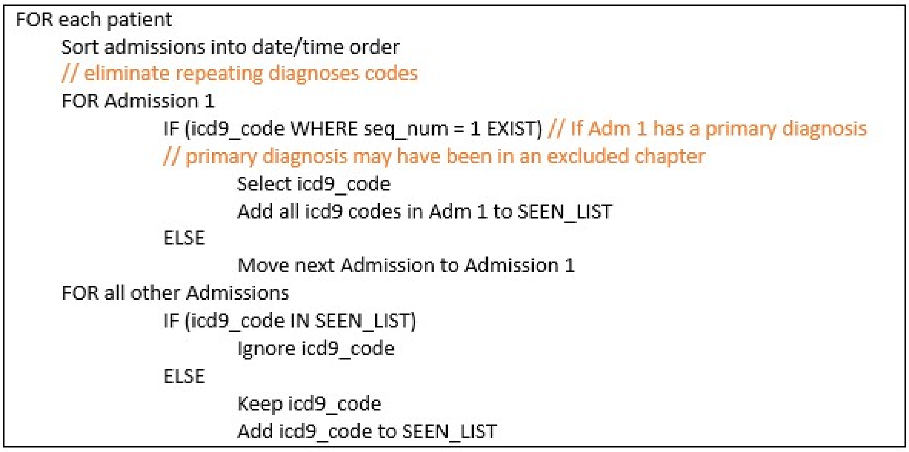
Patient 2	Admission1	Admission 2	Admission 3	Admission 4	Admission 5
Primary diagnosis	**A**	A	**F**	**X**	**E**
Secondary diagnoses	P	**G**	N	**Z**	
	Q	N	A		
	V		V		
	N				

(c)

Patient 3	Admission1	Admission 2
Primary diagnosis	**E**	**F**
Secondary diagnoses	B	
	C	

In Figure 5.3a all secondary diagnoses in the first admission (B, C, J) are discounted. The fact that any subsequent secondary diagnoses were not coded at the first admission (E, F, G) suggests they are newly relevant and should be included. The pseudocode for this processing is shown in Figure 5.4. Figures 5.3b and 5.3c provide two further examples by applying this rule.

Figure 5.4 Pseudo code for deleting repeating diagnoses



```
FOR each patient
    Sort admissions into date/time order
    // eliminate repeating diagnoses codes
    FOR Admission 1
            IF (icd9_code WHERE seq_num = 1 EXIST) // If Adm 1 has a primary diagnosis
            // primary diagnosis may have been in an excluded chapter
                    Select icd9_code
                    Add all icd9 codes in Adm 1 to SEEN_LIST
            ELSE
                    Move next Admission to Admission 1
    FOR all other Admissions
            IF (icd9_code IN SEEN_LIST)
                    Ignore icd9_code
            ELSE
                    Keep icd9_code
                    Add icd9_code to SEEN_LIST
```

Rule 7: Only patients with a single non-repeating diagnosis in each admission are included. Therefore, to allow for a disease 'trajectory', patients must have more than one admission.

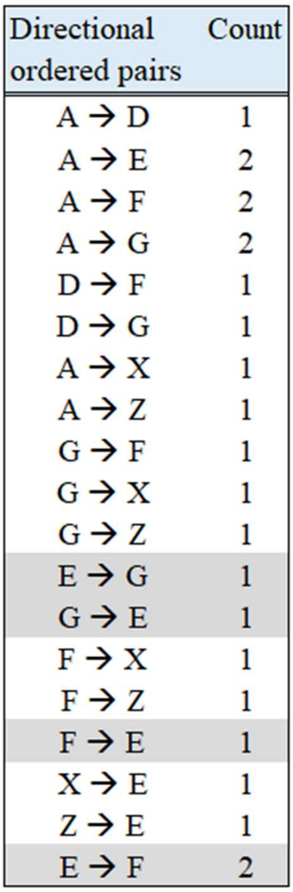
Rule 8: This rule is similar to rule 3, with the exception of all repeating diagnoses are excluded.

Rule 9: Rules 9 and 10 saw a change in how the trajectories were created. The concept of directional ordered pairs was introduced. Here, repeating diagnoses, along with all diagnoses in the first admission are included.

Rule 10: Only primary diagnoses for the first admission are included. All repeating diagnoses are excluded.

When creating directional ordered pairs, sequence strings were built for each patient from the remaining diagnoses and separated into ordered pairs. Table 5.2 presents the list of ordered pairs with their associated counts for the example data in Figure 5.3.

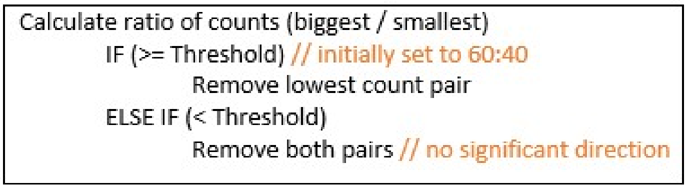
Table 5.2 List of directional ordered pairs



Directional ordered pairs	Count
A → D	1
A → E	2
A → F	2
A → G	2
D → F	1
D → G	1
A → X	1
A → Z	1
G → F	1
G → X	1
G → Z	1
E → G	1
G → E	1
F → X	1
F → Z	1
F → E	1
X → E	1
Z → E	1
E → F	2

A ‘follows’ relationship was used to build the directional ordered pairs. Any two diagnoses, regardless of distance, for the same patient were paired. The focus was on where one disease proceeded another, regardless of diseases in between. Diagnoses within the same admission were not paired, as it would have been unrealistic to assign an ordering to when diseases were contracted. Pairs that appeared in both directions, known as inverse pairs, had their counts compared using a simple ratio. Where one or both pairs did not reach the threshold they were removed. An initial comparison threshold of 60/40 was set (see Figure 5.5). The bottom 20 percent of pair sets was discarded to simplify the model by reducing trace variation. As the frequency of these pair sets was low, it did not impact on the validity of the final trajectory models.

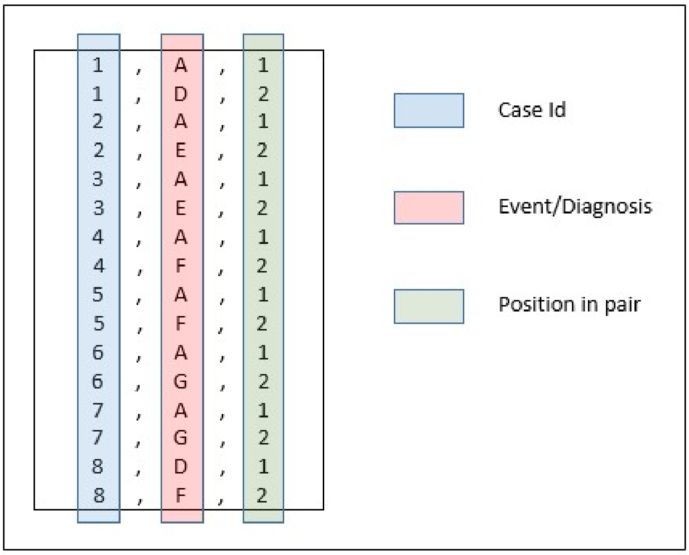
Figure 5.5 Pseudo code for calculating inverse pairs (A→B, B→A)



```
Calculate ratio of counts (biggest / smallest)
        IF (>= Threshold) // initially set to 60:40
                Remove lowest count pair
        ELSE IF (< Threshold)
                Remove both pairs // no significant direction
```

The event log was constructed with the case_id, diagnosis and timestamp. The timestamp was used to determine the position in pair (Figure 5.6).

Figure 5.6 Method to create an event log from the directional ordered pairs



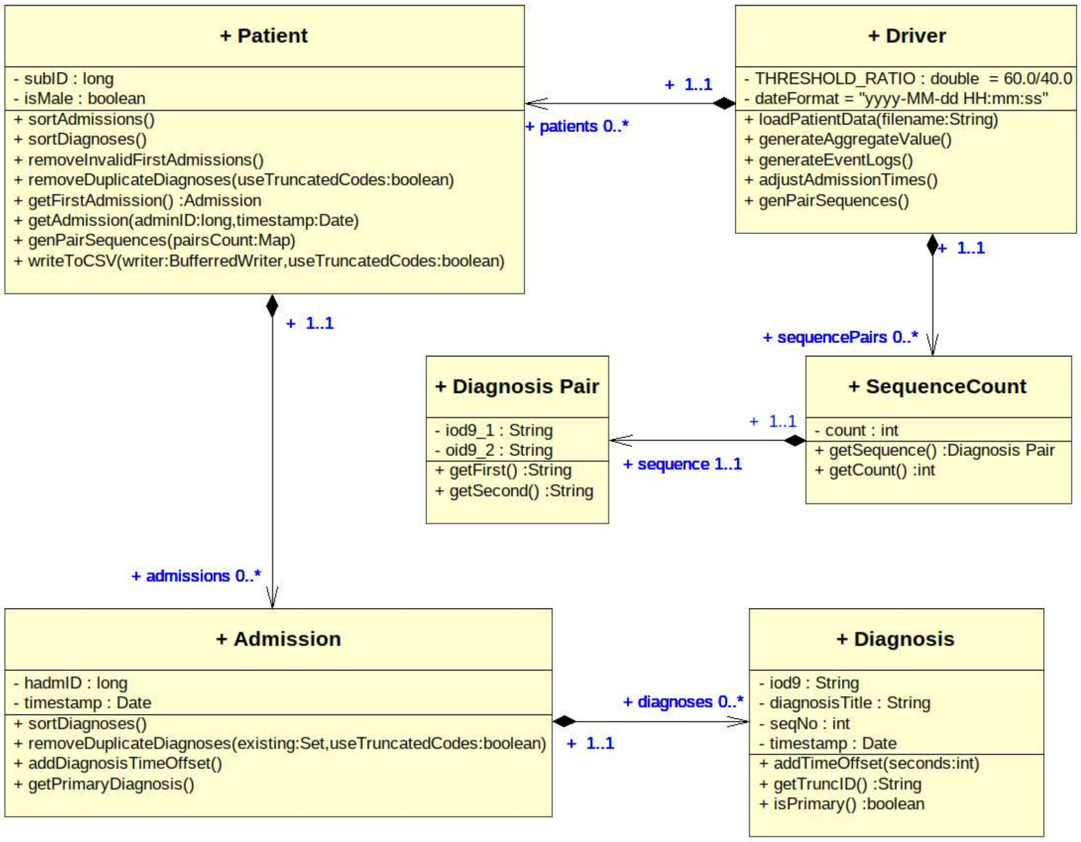
```
1 , A , 1
1 , D , 2
2 , A , 1
2 , E , 2
3 , A , 1
3 , E , 2
4 , A , 1
4 , F , 2
5 , A , 1
5 , F , 2
6 , A , 1
6 , G , 2
7 , A , 1
7 , G , 2
8 , D , 1
8 , F , 2
```

This information shows how the first five directional ordered pairs from Table 5.2 were represented in the event log.

Due to the high level of complexity involved in the data transformations, an external programming resource assisted with the creation of the code. The code for rules 1 to 8 was produced using the Python 3.5 programming language [366] within the Spyder 2.3.9 development environment [367]. Due to the limitations of Python, the code for rules 9 and 10 was created using Java [368] within the Eclipse Integrated development environment (IDE) [369].

In order to design the software for the two experiments a UML class diagram was created using the Quality Software Engineering Environment (QSEE) Superlite v1.1.2 [370]. This class diagram is presented in Figure 5.7.

Figure 5.7 Class diagram for the experiments



For each experiment, the Driver class accepts an input file and determines the sequence in which all other classes and operations are executed.

The final step in this stage was to load the event log files into the process mining tool DISCO to create the disease trajectory models.

5.4 Stage 3: Mining and analysis

The event log files imported into the DISCO process mining tool at the end of the previous stage were used to create the disease trajectory models. As previously stated in Section 3.5, the continuous evolution and rapid change in healthcare processes presents challenges when working with healthcare data. An example of this was evidenced when mapping the ICD-9 codes, used in the MIMIC-III data, to the ICD-10 codes, used in the Jensen model. The mapping between codes proved to be a time consuming task and was carried out using a combination of automated mapping software [371] and manual intervention. This mapping is available in Appendix A.

As stated in Chapter 3, the minimum data items required in an event log is case ID, activity and timestamp. No timestamp information was available for rules 1 and 2, therefore these could not be used for process mining purposes. The event log for Rule 7 contained only three patients, none of whom had a gout diagnosis, therefore, this experiment was also excluded for process mining. The shaded data items, presented in Table 5.3, were included in the seven event logs.

Table 5.3 Event log data items

Rule #	subject_id (CaseID)	hadm_id	timestamp	new timestamp	ICD-9 code (activity)	sequence #	pair ID (CaseID)
3	X	X	X	X	X	X	
4	X	X	X		X	X	
5	X	X	X		X	X	
6	X	X	X		X	X	
8	X	X		X	X	X	
9				X	X		X
10				X	X		X

After the event logs were loaded into DISCO, the data was displayed in the form of a process model (Section 4.1.3.2). All process models were created using the default visualisation settings. When all activities and paths were displayed the models became unreadable due to the high degree of variation in the event log data. Therefore, only the top 4 to 6.6 percent of diseases were displayed for models 3, 4, 5, 6 and 8 (see Appendix B). This was to ensure that the models were simple enough to understand, though retained the maximum number of diagnoses possible. A diagnosis of gout was not included in the top 6.6 percentage of diseases, therefore, separate gout trajectory models were constructed. The first experiment used Rule 9 and the second experiment used Rule 10. The process models for these experiment are presented in this chapter and to be in-line with the Jensen model display 26 diseases. The information presented in Table 5.4 allows for comparison between the 53 disease types displayed in these two models and the Jensen model.

Table 5.4 Disease types included in the three disease trajectory models

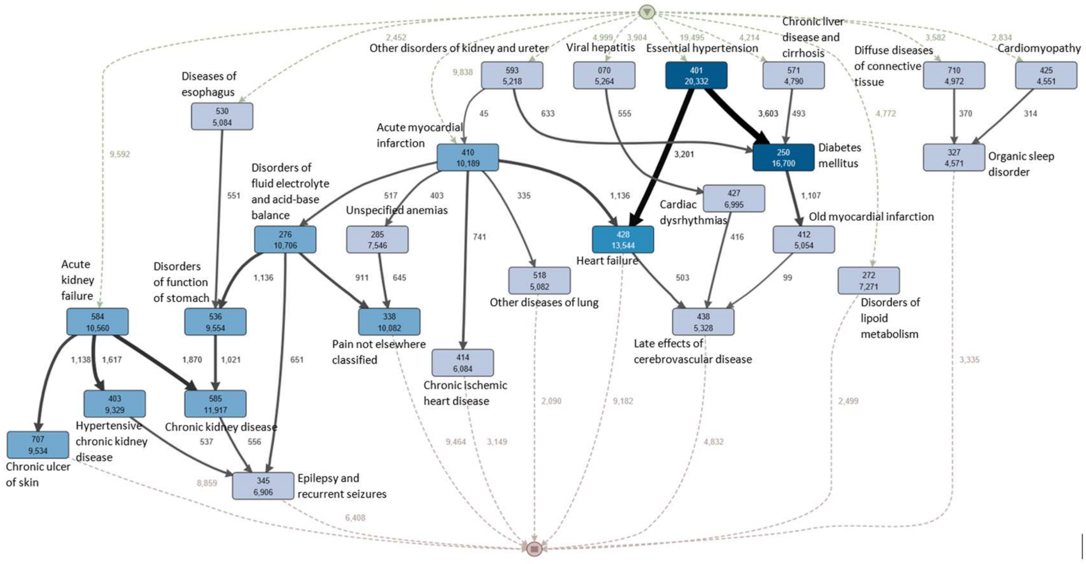
Disease type (ICD-10)	Jensen	Exp 1	Exp 2
Acute myocardial infarction	X	X	X
Anaemias	X	X	X
Chronic ischaemic heart disease	X	X	X
Disorders of fluid, electrolyte and acid-base balance	X	X	X
Heart failure	X	X	X
Atrial fibrillation and flutter (I48)/Cardiac dysrhythmias (I47)	X	X	X
Hypertension (I10)/Hypertensive heart disease (I11)	X	X	X
Diffuse diseases of connective tissue (M32)/Systemic involvement of connective tissue (M35)	X	X	
Sequelae of cerebro-vascular disease	X	X	
Acute kidney failure		X	X
Chronic kidney disease		X	X
Chronic ulcer of skin		X	X
Diabetes mellitus		X	X
Diseases of lung		X	X
Disorders of lipoid metabolism		X	X
Hypertensive chronic kidney disease		X	X
Old myocardial infarction		X	X
Pneumonia	X		X
Acute post-haemorrhagic anaemia	X		
Angina pectoris	X		
Bacterial pneumonia	X		
Bronchiectasis	X		
Caridiac arrest	X		
Complications of heart disease	X		
Cystitis	X		
Gastritis and duodenitis	X		
Gout	X		
Iron deficiency anaemia	X		
Myelodysplastic syndromes	X		
Non-infective gastroenteritis and colitis	X		
Peripheral vascular diseases	X		
Pneumonia due to haemophilus influenza	X		
Rheumatoid arthritis	X		
Volume depletion	X		
Cardiomyopathy		X	
Chronic liver disease and cirrhosis		X	
Diseases of esophagus		X	
Disorders of kidney and ureter		X	
Disorders of function of stomach		X	
Epilepsy and recurrent seizures		X	
Organic sleep disorder		X	
Pain not elsewhere classified		X	
Viral hepatitis		X	
Bacterial infection			X
Candidiasis			X
Conditions of brain			X
Depressive disorder			X
Diseases of endocardium			X
Disorders of urethra and urinary tract			X
Hypotension			X
Intestinal infections due to other organisms			X
Purpura and other hemorrhagic conditions			X
Septicemia			X

The two main disease trajectory models created by generating directional ordered pairs of diagnoses using Rules 9 and 10 are presented in the two sub-sections below.

5.4.1 Experiment 1: Disease trajectory model following Jensen's rules

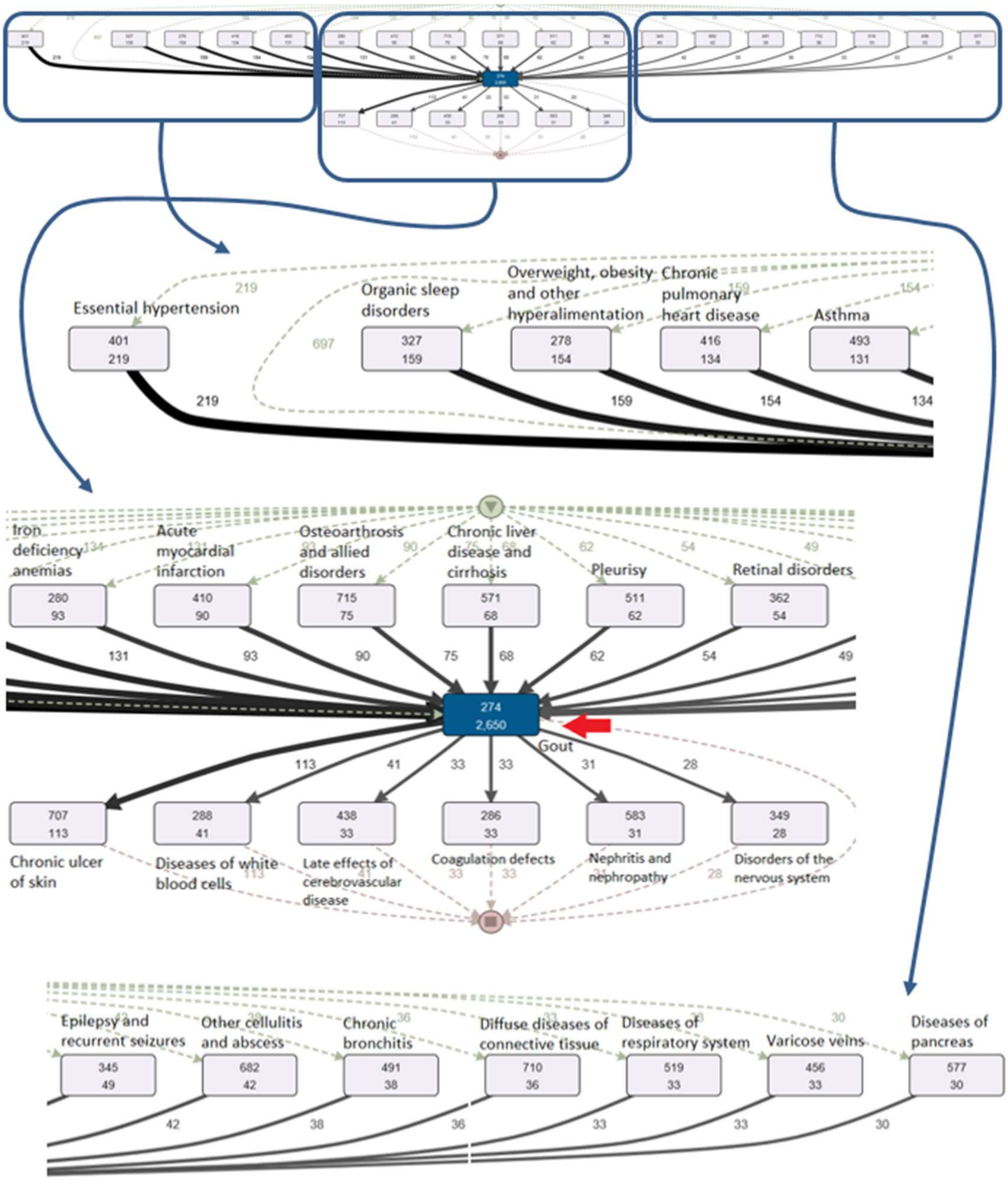
Figure 5.8 presents the CVD trajectory model with level three ICD-9 codes and directional ordered pairs of diagnoses. In-line with the Jensen model, only the top 26 diagnosis types are displayed. The rules for this experiment align most closely with those of Jensen's.

Figure 5.8 Cardiovascular disease trajectory model for Experiment 1



When compared to the Jensen model, nine of the diagnosis types also existed in the Jensen model, these are discussed later in the evaluation stage. There is no gout diagnosis on this model. It is not until the number of diagnoses displayed on the model is increased to 12.3 percent (n=67) that gout becomes visible. Figure 5.9 presents the gout trajectory model.

Figure 5.9 CVD trajectory model for Experiment 1 filtered on gout


Essential hypertension
401
219
Organic sleep disorders
327
159
Overweight, obesity and other hyperalimentation
278
154
Chronic pulmonary heart disease
416
134
Asthma
493
131
219
697
219
159
154
134
Iron deficiency anemias
280
93
Acute myocardial infarction
410
90
Osteoarthrosis and allied disorders
715
75
Chronic liver disease and cirrhosis
571
68
Pleurisy
511
62
Retinal disorders
362
54
131
93
90
75
68
62
54
49
274
2,650
Gout
113
41
33
33
31
28
Chronic ulcer of skin
707
113
Diseases of white blood cells
288
41
Late effects of cerebrovascular disease
438
33
Coagulation defects
286
33
Nephritis and nephropathy
583
31
Disorders of the nervous system
349
28
Epilepsy and recurrent seizures
345
49
Other cellulitis and abscess
682
42
Chronic bronchitis
491
38
Diffuse diseases of connective tissue
710
36
Diseases of respiratory system
519
33
Varicose veins
456
33
Diseases of pancreas
577
30
42
38
36
33
33
30

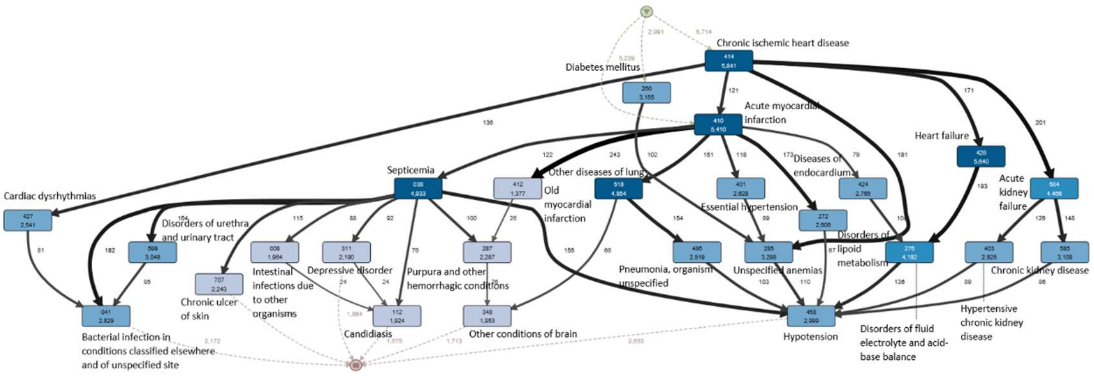
When comparing this model against the Jensen model, there are an additional three diagnosis types in common, these are: gout, iron deficiency anaemia and bronchiectasis.

5.4.2 Experiment 2: Refined disease trajectory model

The final disease trajectory model in this series was produced using Rule 10. After discussions with the clinical domain experts and with Jensen, the decision was made to extend the rules of experiment 1 to exclude repeating diagnoses for patients. This is because the majority of diseases within the CVD cluster are chronic in nature and as such, the assumption was made that any subsequent repeat diagnoses for patients were most likely not new occurrences of the disease, rather a re-recording in the patient's EHR.

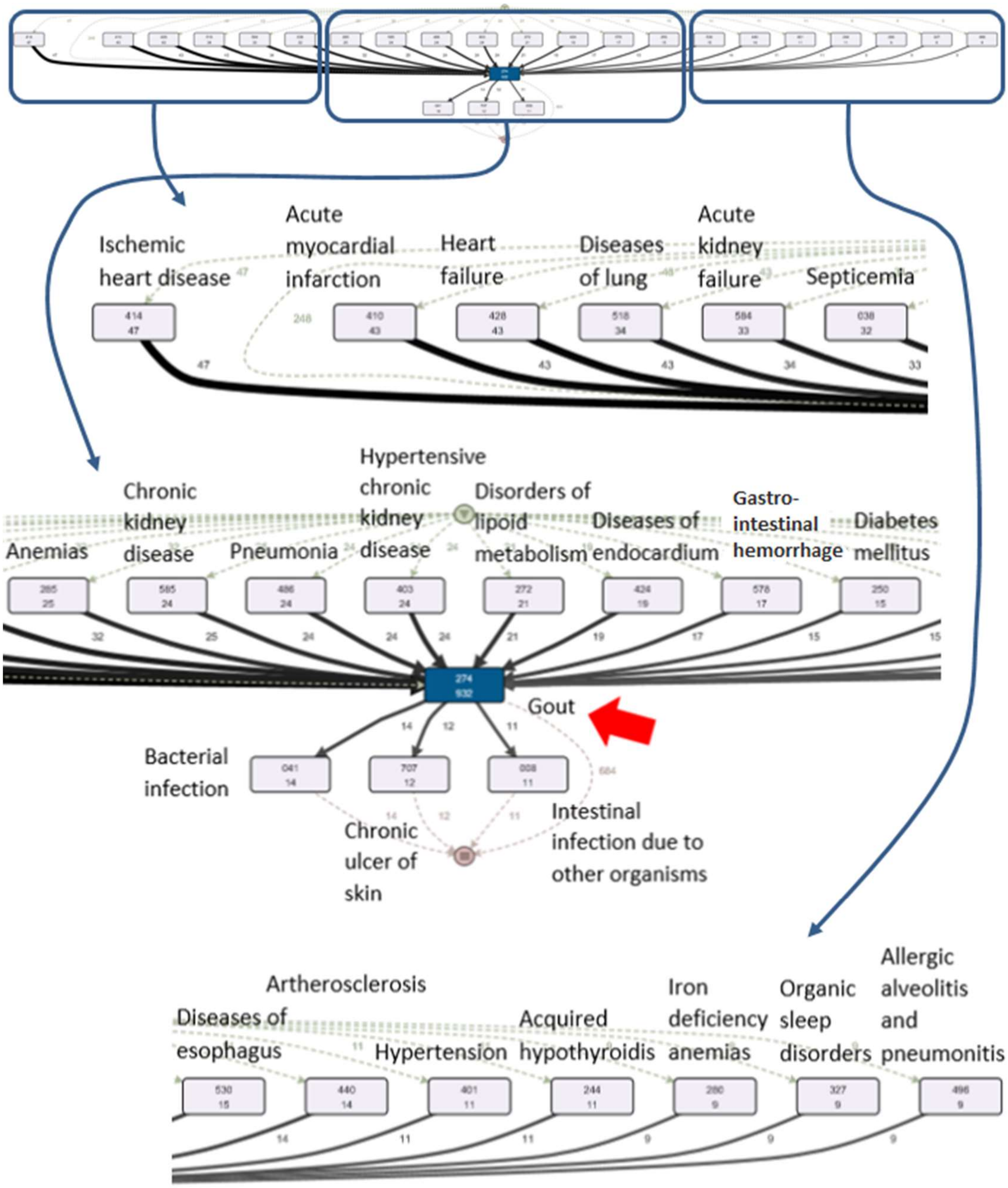
Figure 5.10 presents this CVD trajectory model where directional ordered pairs of diagnoses were used to construct the model, ICD-9 codes are aggregated to level three and all repeating patient diagnoses have been excluded. Again, in-line with the Jensen model and with Experiment 1, only the top 26 diagnosis types are displayed.

Figure 5.10 Cardiovascular disease trajectory model for Experiment 2



When this model is compared to the Jensen model, eight diagnosis types are common (see Table 5.4). Although, when this model is compared against the model created in Experiment 1, 15 of the diagnosis types were common. These commonalities are discussed during the evaluation stage.

Figure 5.11 CVD trajectory model for Experiment 2 filtered on gout

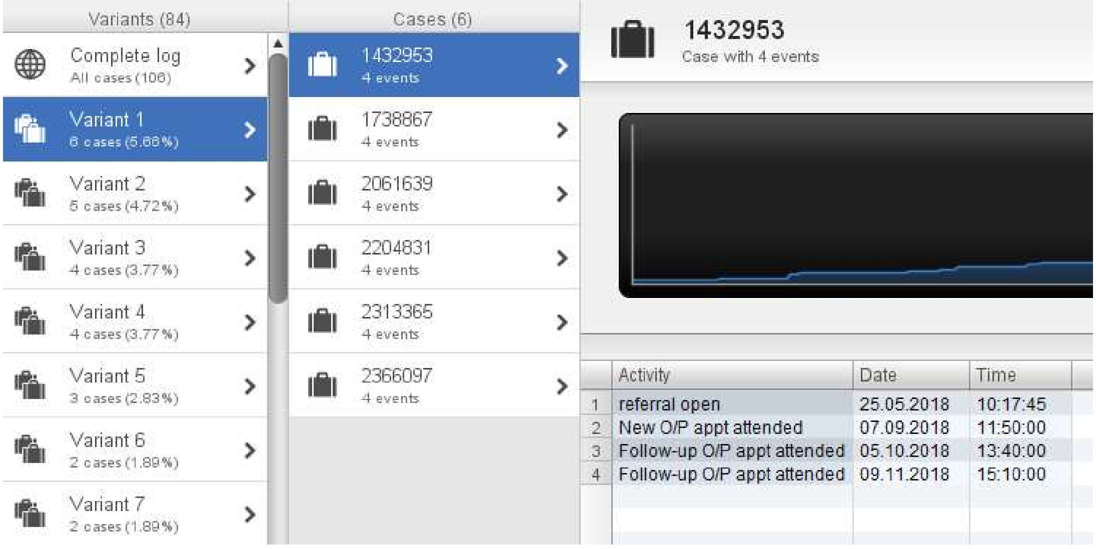


In the refined model gout did not present in the top 26 diagnosis types. It was not until 11.6 percent (n=63) of diagnosis types were displayed that gout became visible within the model. Figure 5.11 presents the gout trajectory model. When comparing this model against the model from Experiment 1, there are seven diagnosis types in common these are: gout; acute myocardial infarction; diseases of the lung/chronic pulmonary heart disease; hypertension; iron deficiency anaemia; chronic ulcer of skin; and organic sleep disorders.

5.5 Stage 4: Evaluation

The aim of this study was to establish whether process mining techniques could be used to create disease trajectory models using the MIMIC-III data. First, a series of ten rules was created based on those of Jensen. Seven event logs were then created using the MIMIC-III data. These event logs contained data related to patients diagnosed with CVD. Disease trajectory models were discovered using the process mining tool DISCO. As DISCO uses the DISCO Miner algorithm (see Section 3.4.5), all discovered models accurately represent the event log data. Validation of the models was performed manually by carrying out a number of checks for each model. Event log data was compared using the visualisations provided within the tool. Basic checks such as the number of cases in each model were carried out using the Map or Statistics view. Figure 5.12 demonstrates how the data was viewed at case level with the related variant, case and activity-level statistics. A sample of data for each model was randomly selected and checked using this method.

Figure 5. 12 Variant, case and activity level view of the data



Variants (84)

Complete log
All cases (106)

Variant 1
6 cases (5.66%)

Variant 2
5 cases (4.72%)

Variant 3
4 cases (3.77%)

Variant 4
4 cases (3.77%)

Variant 5
3 cases (2.83%)

Variant 6
2 cases (1.89%)

Variant 7
2 cases (1.89%)

Cases (6)

1432953
4 events

1738867
4 events

2061639
4 events

2204831
4 events

2313365
4 events

2366097
4 events

1432953
Case with 4 events

	Activity	Date	Time
1	referral open	25.05.2018	10:17:45
2	New O/P appt attended	07.09.2018	11:50:00
3	Follow-up O/P appt attended	05.10.2018	13:40:00
4	Follow-up O/P appt attended	09.11.2018	15:10:00

The results were further validated by comparing them to the Jensen results. The features of Experiment 1 most closely replicated those used by Jensen and refinements were made in Experiment 2. In this section, these two disease trajectory models are interpreted and discussed. Verification is carried out by comparing the first model to the Jensen model. A clinical expert from within the field of cardiovascular medicine evaluated the correctness of the findings presented in Section 5.4. Finally, the strengths and weaknesses of using a process mining approach to create disease trajectory models are discussed.

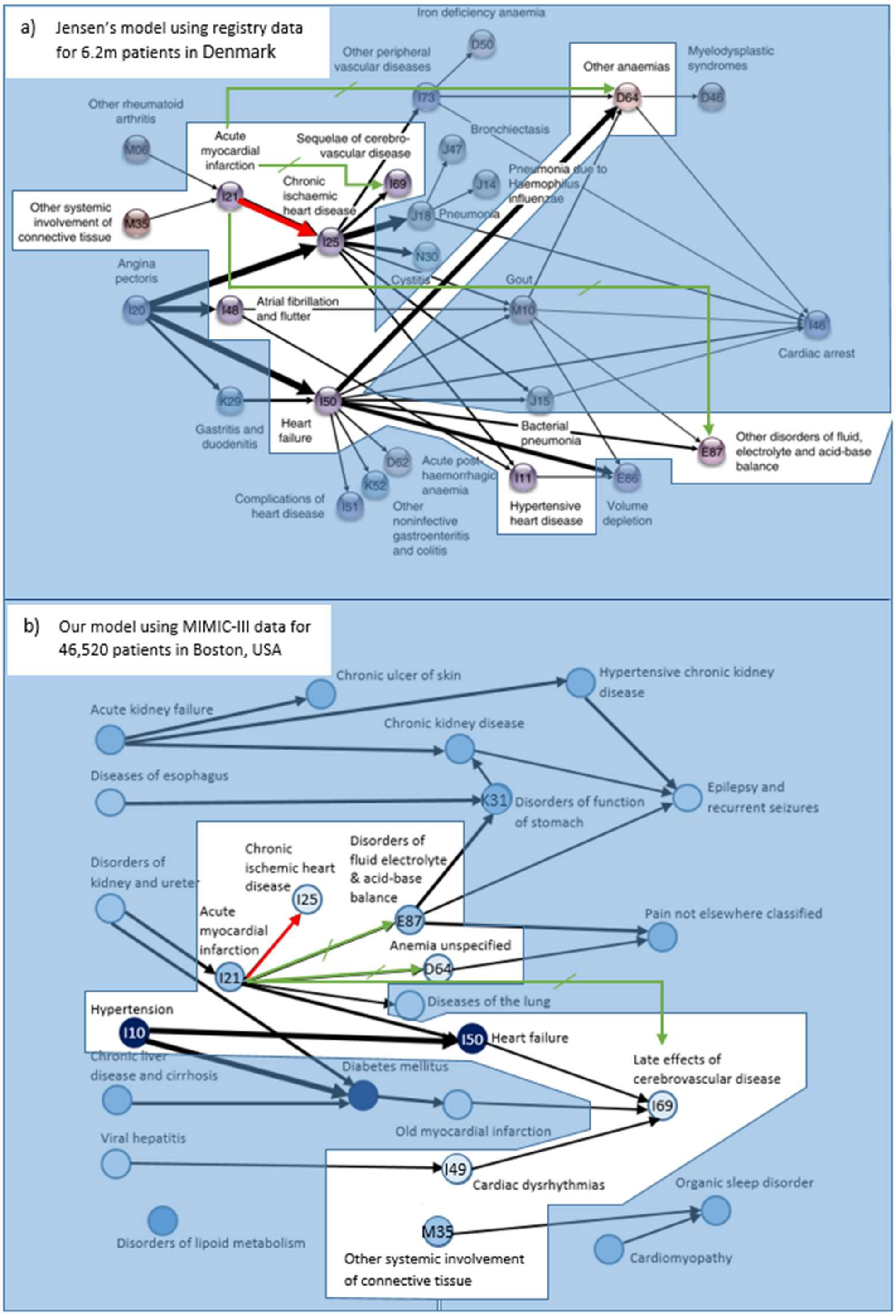
5.5.1 Discussion and comparison of results

In this section, the three disease trajectory models are compared and discussed. First, the Jensen model is compared to the model from Experiment 1. Second, the model from Experiment 1, is compared to the refined model from Experiment 2.

1) *Comparison between Jensen's model and the model from Experiment 1*

The information presented in Figure 5.13 shows the commonalities between the two models. Associations between disease types are represented by directional arcs, where the thickness of the arc indicates the frequency of diseases. A red arrow represents an identical relationship in both models and a green arrow with a bar represents a 'follows' relationship in both models. ICD-10 disease codes are displayed inside the circular nodes. For comparison purposes, all ICD-9 codes in Rule 9 have been converted to ICD-10 as the Jensen model (see Appendix A for mapping).

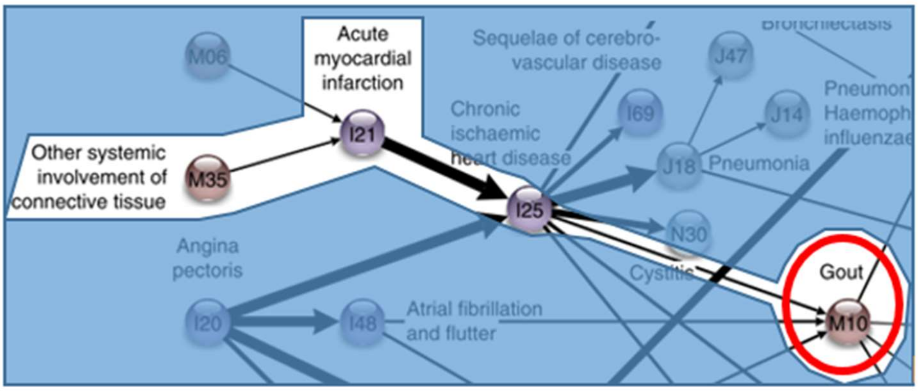
Figure 5.13 Comparison of the Jensen model and the model from Experiment 1


a) Jensen's model using registry data for 6.2m patients in Denmark
Iron deficiency anaemia
D50
Other peripheral vascular diseases
I73
Other anaemias
D64
Myelodysplastic syndromes
D46
Other rheumatoid arthritis
M06
Acute myocardial infarction
I21
Sequelae of cerebrovascular disease
I69
Chronic ischaemic heart disease
I25
Bronchiectasis
J47
J14
Pneumonia due to Haemophilus influenzae
J18
Pneumonia
N30
Cystitis
Other systemic involvement of connective tissue
M35
Angina pectoris
I20
I48
Atrial fibrillation and flutter
Gout
M10
I46
Cardiac arrest
K29
Gastritis and duodenitis
I50
Heart failure
J15
Bacterial pneumonia
E87
Other disorders of fluid, electrolyte and acid-base balance
D62
Acute post-haemorrhagic anaemia
K52
I51
Complications of heart disease
Other noninfective gastroenteritis and colitis
I11
Hypertensive heart disease
E86
Volume depletion
b) Our model using MIMIC-III data for 46,520 patients in Boston, USA
Chronic ulcer of skin
Hypertensive chronic kidney disease
Acute kidney failure
Chronic kidney disease
Diseases of esophagus
K31
Disorders of function of stomach
Epilepsy and recurrent seizures
Disorders of kidney and ureter
Chronic ischemic heart disease
I25
Disorders of fluid electrolyte & acid-base balance
E87
Acute myocardial infarction
I21
Anemia unspecified
D64
Pain not elsewhere classified
Diseases of the lung
Hypertension
I10
I50
Heart failure
Late effects of cerebrovascular disease
Chronic liver disease and cirrhosis
Diabetes mellitus
I69
Old myocardial infarction
Viral hepatitis
I49
Cardiac dysrhythmias
Organic sleep disorder
M35
Disorders of lipoid metabolism
Other systemic involvement of connective tissue
Cardiomyopathy

Nine disease types are common in both models, these include: acute myocardial infarction; chronic ischemic heart disease; heart failure; hypertension; sequelae of cerebrovascular disease; anaemia; and disorders of fluid electrolyte, acid-base balance; atrial fibrillation and flutter; and other systemic involvement of connective tissue. An exact association exists from acute myocardial infarction to chronic ischaemic heart disease. Three similar ‘follows’ relationships exist from acute myocardial infarction to: disorders of fluid electrolyte and acid-base balance; anaemia; and late effects of cerebrovascular disease. Related diseases between the two models include gastritis and duodenitis in model a), and disorders of function of stomach in model b). Here, there is a reversed ‘follows’ relationship with disorders of fluid electrolyte and acid-base balance. From a clinical perspective, the direction of the flow in model b) is more logical, as medication taken after an acute myocardial infarction can often lead to disorders of fluid electrolyte and acid-base balance, and may also be associated with disorders of function of the stomach. It must be remembered, that when creating directionality between the ordered pairs a threshold value of 40/60 was applied. This may contribute towards the reason that some relationships between similar diseases appear in opposite directions on both models.

Jensen’s model is used in Figure 5.14 to highlight the two similarities found between the models in relation to gout. In Experiment 1, systemic involvement of connective tissue and acute myocardial infarction are both direct predecessors to gout.

Figure 5.14 Comparison of gout associations for Experiment 1

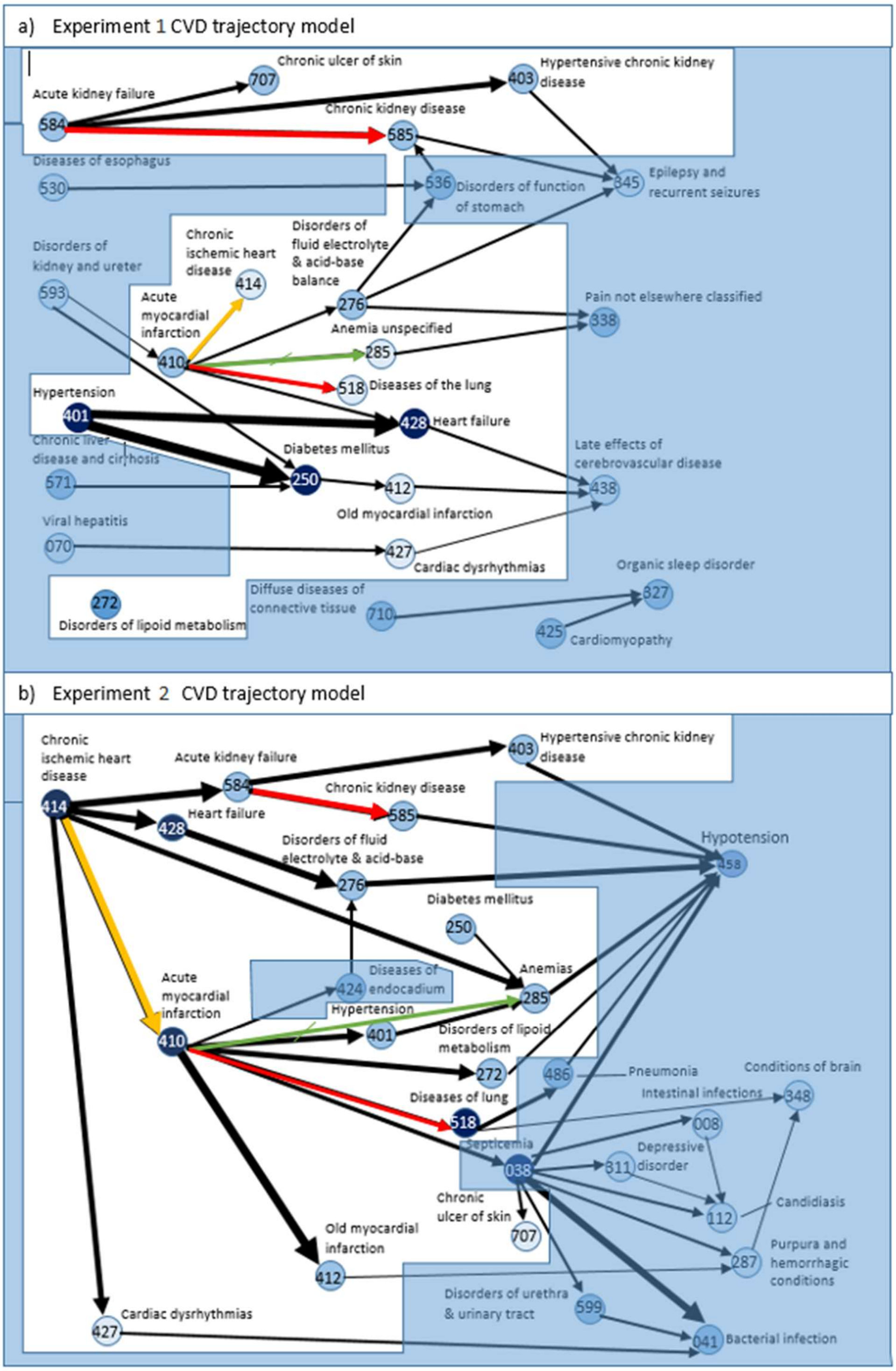

Bronchiectasis
M06
Acute myocardial infarction
Sequelae of cerebro-vascular disease
J47
I21
Chronic ischaemic heart disease
I69
J14
Pneumon
Haemoph
influenzae
Other systemic involvement of connective tissue
M35
J18
Pneumonia
I25
N30
Angina pectoris
Cystitis
Gout
I20
I48
Atrial fibrillation and flutter
M10

2) *Comparison between models from experiments 1 and 2*

The information presented in Figure 5.14 shows the commonalities between the models created in experiments 1 and 2. In addition to the features used in Experiment

1, the refined model in Experiment 2 excludes secondary diagnoses in the first admission and repeating diagnosis codes for patients. Here, the assumption was made that the first time a disease is seen in the EHR for a patient, is the first time it has appeared for that patient. This assumption relies on all previous diagnoses for that patient being recorded at the first admission.

Figure 5.15 Comparison of the models created in experiments 1 and 2


a) Experiment 1 CVD trajectory model
Acute kidney failure
584
Chronic ulcer of skin
707
Hypertensive chronic kidney disease
403
Chronic kidney disease
585
Diseases of esophagus
530
536
Disorders of function of stomach
345
Epilepsy and recurrent seizures
Disorders of kidney and ureter
593
Chronic ischemic heart disease
414
Disorders of fluid electrolyte & acid-base balance
276
Acute myocardial infarction
410
Anemia unspecified
285
Pain not elsewhere classified
338
518
Diseases of the lung
Hypertension
401
428
Heart failure
Chronic liver disease and cirrhosis
571
Diabetes mellitus
250
412
Old myocardial infarction
Late effects of cerebrovascular disease
438
Viral hepatitis
070
427
Cardiac dysrhythmias
Organic sleep disorder
327
272
Disorders of lipoid metabolism
Diffuse diseases of connective tissue
710
425
Cardiomyopathy
b) Experiment 2 CVD trajectory model
Chronic ischemic heart disease
414
Acute kidney failure
584
403
Hypertensive chronic kidney disease
Chronic kidney disease
585
Heart failure
428
Hypotension
458
Disorders of fluid electrolyte & acid-base
276
Diabetes mellitus
250
Anemias
285
Diseases of endocadium
424
Hypertension
401
Acute myocardial infarction
410
Disorders of lipoid metabolism
272
486
Pneumonia
Conditions of brain
Intestinal infections
348
008
Diseases of lung
518
Septicemia
038
Depressive disorder
311
112
Candidiasis
Chronic ulcer of skin
707
Old myocardial infarction
412
287
Purpura and hemorrhagic conditions
Disorders of urethra & urinary tract
599
Cardiac dysrhythmias
427
041
Bacterial infection

When comparing these models, 15 disease types are common in both models. Two exact associations exist between acute kidney failure and chronic kidney disease, and acute myocardial infarction and diseases of the lung. A similar 'follows' association exists between acute myocardial infarction and anaemia. Finally, there is a reverse association between chronic ischemic heart disease and acute myocardial infarction. Additional disease types present in model b) include: pneumonia; bacterial infection; candidiasis; conditions of brain; depressive disorder; diseases of endocardium; disorders of urethra and urinary tract; hypotension; intestinal infections; purpura and other hemorrhagic conditions; and septicaemia. As opposed to many of the 26 disease types present in model a) but not in model b), many of these 10 are acute rather than chronic conditions. One explanation for this difference, may be that when repeated patient diagnoses are excluded from the analysis (re-recordings of chronic diseases), it makes way for more frequently recorded, non-long term conditions such as the infections present in model b).

When comparing the gout trajectory models, there are seven diseases in common between the two experiments. It is interesting to note that when repeating patient diagnoses are excluded, the model includes more infection-related diseases such as septicemia and pneumonia. NSAIDs, corticosteroids and Febuxostat, a urate-lowering therapy, are all pharmaceutical treatments employed in patients suffering with gout [200]. A recent review of the gout treatment guidelines, stated that patients with CVD should avoid NSAIDs due to an observed increase in the cardiovascular event rate identified in large cohort studies. This increase was particularly noticeable with myocardial infarctions, strokes, and heart failures [197]. In addition, the chronic inflammation due to urate deposition in gout, which although principally apparent in joints occurs throughout the body is, itself linked to higher rates of CVD and chronic kidney disease [198]. Finally, gout is known to be associated with comorbidities such as atrial fibrillation, obstructive sleep apnoea, osteoporosis and venous thromboembolism. Hence, there are several reasons why in cohort studies, gout may co-locate with CVD and its consequences.

The method for mining and mapping of disease trajectories using process mining tools may be of interest to healthcare researchers, as it enables a more rapid modelling, made possible by the transformation of diagnosis pairs into an event log. However, one of the most valuable pieces of information prior to using any method, would be to know whether a diagnosis was pre-existing for a patient. A challenge when working with healthcare data is that the data is only a partial view on reality (see Section 3.5.1.2). In this study, to overcome the issue of uncertainty, caused by potentially missing diagnoses, a ‘window of interest’ was created. This was done by retaining only the primary diagnosis for the first admission, as it is unknown whether the secondary diagnoses in the first admission repeated from a previous time prior to the window. The assumption was made that, apart from secondary diagnoses in the first admission, the first time a diagnosis was recorded in the data, was the first time it occurred in the patient. These unused secondary diagnoses in the first admission were retained for comparison against all diagnoses in future admissions. If a match was made the diagnosis was discarded. By doing this, it can be more safely assumed that any valid diagnoses happened within the window of interest. Compared with the method adopted by Jensen, this leads to a more conservative estimate with the data. For example, consider a trajectory of A > B > C. With Jensen’s method this translates to including [A] > [B] > [C]; [A, B] > [C]; [A] > [B, C], where the square brackets denote diseases recorded within the same admission. Using the method applied to the MIMIC-III data, the same trajectory would translate only to [A] > [B] > [C]. Jensen remarked [75] that as their method was already conservative by having the requirement where patients must follow a trajectory of four diagnoses in that certain order, they therefore did not want to place further restrictions on the data. After careful consideration with the clinical experts, it was decided that including diagnoses from the same admission would introduce an incorrect ordering between secondary diagnoses within an admission. This being the case, we did not further restrict the data by considering time between admissions or a minimum length for trajectories. It can therefore be said that disease trajectories alone, from a general hospital population do not necessarily uncover causality, rather they reveal a mixture of biology, hospital or health care system practice and organisation of care.

There are a number of limitations to this work. The first relates to the differences between primary and secondary care data. As gout is predominantly diagnosed and treated in primary care by GPs, unless it is the primary reason for attendance, the diagnosis is often not recorded in secondary care data. Whereas, a diagnosis made in secondary care is more likely to be reliably recorded in the primary care record, unfortunately the reverse does not apply.

Coded data is recorded primarily for billing and administrative purposes rather than quality of care and monitoring. This is a second limitation, as it may introduce bias to diagnoses yielding a higher revenue with certain other diagnoses going unrecorded, for example QOF.

A third limitation relates to potential discrepancies in the mapping of diagnosis codes between the two versions of the ICD system. This is due to not all diseases having a one to one relationship.

Creating a 'window of interest' assumes all existing comorbidities are recorded for a patient at the first admission, which is not always the case. The final limitation relates to the assumption that any valid diagnoses happened within the window of interest. This requires all existing comorbidities to have been recorded for a patient at the first admission, which is clearly not always the case.

5.5.2 Clinical evaluation

From a clinical perspective, the vast majority of the diseases in both models follow a logical progression. Some that would benefit from further exploration include epilepsy and recurrent seizures and the link between viral hepatitis and cardiac dysrhythmias. The diseases in Experiment 1 can be described by following an illustrative example for the separate journeys of John and Jane.

In 2016 John suffered an acute myocardial infarction, shortly followed by lung cancer and anaemia. The high amounts of medication John was taking after the myocardial infarction, including ACE inhibitors and asprin, lead to disorders of fluid electrolyte and acid-base balance. Eventually, John was diagnosed with a stomach ulcer and chronic kidney disease which resulted in chronic ulcers of the skin due to peripheral edema and ankle swelling. Jane is another patient who has been diagnosed with many of the diseases from the cardiovascular cluster. Jane suffered from essential hypertension and chronic liver disease for some time before presenting with diabetes mellitus and heart failure. On further investigation for her diabetes, Jane was diagnosed with an old myocardial infarction. However, with close monitoring by Jane's GP and changes to her lifestyle, Jane managed to live a fairly normal life until many years later when she began to experience the late effects of cerebrovascular disease.

When creating the model from Experiment 2, a large number of repeated diagnoses were removed from the data, resulting in a higher level of sensitivity. This level of sensitivity allowed for new relationships to be identified, which were previously swamped by large amounts of unnecessary data. When exploring the differences between the models from the two experiments, the appearance of hypotension, pneumonia and depressive disorder all make sense from a clinical perspective. Myocardial infarction is a form of ischemic heart disease and therefore implies the presence of chronic ischemic heart disease, which may or may not be symptomatic. When considering the reversal of the association between these two conditions in Figure 5.14, it is likely to be explained by patients who were admitted with chest pain but did not fulfil the criteria for an acute myocardial infarction, though who had chronic ischemic heart disease. The data will contain many primary diagnoses for chronic ischemic heart disease, where patients attended secondary care with chest pain but had normal blood tests (troponin), and a normal electrocardiogram (ECG). Hence secondary care diagnoses of chronic ischaemic heart disease are closely linked or overlap with acute myocardial infarction making true directionality difficult to establish.

Validation of the two disease trajectory models along with our findings was performed by Dr Klaus Witte, a senior lecturer and consultant cardiologist at the University of Leeds. The two disease trajectory models were assessed for their clinical plausibility. The evaluation showed that both models made sense from a clinical perspective, though the improved model, created in the second experiment, included more of the diseases that Dr Witte would have expected to be in the model.

Gout is associated with CVD, probably due to a combination of its inflammatory component and the use of certain medication including NSAIDS, since several of the medications are used for heart failure, hypertension, and therefore post myocardial infarction state, can cause or exacerbate gout. Directionality will depend upon the patient, the setting and the relevance that different physician groups place on it.

5.6 Impact and future work

The impact of the work in this section includes the publication of three peer-reviewed conference proceedings and a poster. The work was first presented at an international conference [360]. Later, work up to and including the data transformation stage in Section 5.3 was used by a member of the University of Leeds process mining research group, Guntur Kusuma. Kusuma extended the work by using a binomial test to reduce the complexity of the event data by measuring the correlation of the identified pairs of diagnostic codes. In addition, Kusuma built a pipeline in order to standardise much of the programming. Resulting from this work, two publications were co-authored. The first, a feasibility study that uses process mining techniques to create disease trajectories [361] and the second, a literature review on process mining of disease trajectories [362]. Following on from [361], the work was further developed and refined by Kusuma et al. in, leading to the publication on process mining of disease trajectories using the MIMIC-III dataset [329]. The combination of process mining and the statistical analyses to identify the disease trajectories of event data using MIMIC-III has shown a potential for application using a larger dataset such as a nation-wide EHR, with the opportunity of international comparison.

Areas for future work may include the exploration by a clinical researcher into anomalies identified during the evaluation stage. Examples of anomalies such as the inclusion of epilepsy and recurrent seizures and the link between viral hepatitis and cardiac dysrhythmias. Other areas may include the creation of disease trajectory models using different datasets for other chronic or acute clusters of diseases to assess the reproducibility of the method. The use of different coverage and in different geographical areas. Other types of data could be incorporated, for example, biomarkers and medication. This would allow for the robustness of the method to be assessed and could also provide valuable information and insights. The process of excluding repeat diagnosis codes could be further improved by classifying them as chronic or acute and applying a timescale cut-off or by considering concomitant medication use. Only repeated chronic diagnosis codes would then be excluded from the event log, as it is likely that acute diseases which were repeated from a previous admission would be new occurrences. As mentioned in Section 5.5.1, a limitation when using secondary care data, is that gout is mainly recorded in primary care systems. Therefore, it may be interesting to perform a similar analysis using linked data from primary and secondary care to establish whether gout becomes more dominant within the cardiovascular cluster of diseases. Within the field of computational linguistics, temporal directional pairs of diagnosis codes may be described as bigrams [372]. Bigrams in computational linguistics is a sequence of two adjacent elements. The use of n-grams and computational linguistics to create disease trajectory models may be an alternative method to explore.

5.7 Summary

The work in this chapter has pioneered the use of process mining techniques for disease trajectory modelling. In this study, process mining techniques were used to create a series of disease trajectory models using the MIMIC-III data. Event logs were created in order to discover seven disease trajectory models using the process mining tool DISCO. The rules for the model generated in the first experiment closely replicated those used by Jensen. Similarities and differences between the models were identified and discussed and all findings validated by a consultant cardiologist at the University of Leeds. Gout was found to be, though not central, an important disease within the cardiovascular cluster of diseases. Refinements were made to the method and a new disease trajectory model was created and compared against our first model. Similarities and differences between the two models were discussed and validated with a consultant cardiologist, where it was considered more clinically relevant to a general population, due to the filtering out of unnecessary volume.

To answer the first research question, data for CVD patients was extracted from the MIMIC-III database. Ten data transformation rules were applied in order to perform two experiments. Two event logs were created containing directional ordered pairs of diseases. To answer the second research question, two disease trajectory models were created in the DISCO process mining tool using the two event logs. The first model followed Jensen's rules and the second was a refinement, where repeating diagnoses for a patient were excluded. To answer the third research question, the results from first disease trajectory model were compared against those from the Jensen model. Nine of the diagnosis types were common to both models. Though gout was not a central diagnosis within the CVD cluster, it was in the top 12 percent of diseases. To answer the fourth research question, after clinical evaluation, it was found that the refined model had a higher level of sensitivity due to the absence of repeated patient diagnoses in the event log data. This resulted in the identification of new relationships. Finally to answer the fifth research question, a summarised list of the main strengths and weaknesses of using a process mining approach to construct disease trajectory models is presented below. In Section 3.5.1 a number of challenges were identified, specifically for process mining healthcare data. Throughout this chapter specific examples of these challenges and how they have been approached are explained.

Main strengths of using a process mining approach to construct disease trajectory models:

1. Automatic creation of professionally drawn disease trajectory models using event log data and standard process mining tools.
2. Automatic calculation and representation of durations between diseases using event log data and standard process mining tools.
3. Flexibility, using slider bars to set the desired level of detail (number of diseases and connections). A useful feature when working with data that has a high degree of variation.
4. Standard process mining tools can be used to perform conformance checking.
5. The creation of an event log is made easier by the hierarchical structure of the clinical coding systems, as it can simplify the data selection code.
6. Comprehensive online documentation for the MIMIC-III data.

Main weaknesses of using a process mining approach to construct disease trajectory models:

1. Event log data for disease trajectory models requires a high amount of pre-processing.
2. Missing diagnoses in healthcare data.
3. Potential bias in the data due to the primary purpose of data collection.
4. Clinical domain expertise required throughout. Availability is often limited due to increasing pressures within the NHS.
5. Data collected from two healthcare information systems with different identifiers.
6. Diseases diagnosed in primary care are often missing from hospital data.

The main contribution from this study is intended to advance knowledge in the field of process and data science. The findings suggest that the method for data mining and mapping of disease trajectories using process mining tools to be of interest to healthcare researchers, as it enables a more rapid modeling and visualisation, made possible by the transformation of diagnosis pairs into an event log. In addition, the results may be further investigated by clinical researchers and used in combination with existing research, to help inform clinical professionals in the field of CVD and gout. This work has highlighted many of the complexities involved in creating disease trajectories from EHRs but has demonstrated that such an approach is feasible.

Disease trajectories are an excellent introduction for working with care pathways and in many ways are a simplification. They can be used, along with other methods, as a way to identify early indicators for associated diseases and generate aetiological hypothesis. The current literature recommends no standard for creating disease trajectory models. Process mining methods have proven to be successful for creating these models and in contrast to other methods stated in Chapter 3, can relatively quickly define disease trajectories.

Though often used at the patient level, by using these well-established process mining software tools, the drawing of disease trajectory models is made much easier. There is no equivalent way. However, the drawing is only part of the problem. Due to the complexity of the transformation rules and the challenges of working with healthcare data, pre-processing of the event log data requires significant programming skill and effort. In order to work towards a fully automated solution, Kusuma [329] has taken this work and standardised much of the programming effort. As a result, the creation of disease trajectory models is now accessible to a larger number of people. By using the MIMIC-III data, the work in this chapter has demonstrated that such an approach is possible.

Chapter 6
Process mining MyPathway data to identify patterns of care

6.1 Introduction

Service providers in industry sectors such as healthcare, education, finance and tourism are increasingly using mobile software platforms to communicate with their patients and clients. This form of digital communication has the potential to improve processes, resulting in a higher level of efficiency and quality, often at a reduced cost. The use of healthcare applications have been associated with improved health outcomes, both self-reported and objectively measured [373]. In-line with this increased use of mobile technology, there is a mounting dependence on Patient Reported Outcome Measure (Section 2.3.2) data [374] as an evidence-based method for objective health assessment.

The work presented in this chapter was carried out whilst working as part of the development team within Advanced Digital Innovation UK Limited (Section 1.4.1). ADI are a privately owned software company and the developers of the successful mobile health application, MyPathway [375]. The MyPathway application is used in various healthcare settings (see Section 6.3.1 for full list), though for this study, data was collected from MSK patients attending Sheffield Teaching Hospitals NHS Foundation Trust (see Section 6.3.2 for details of how patients are managed). The AIM-FORE project [376] was a collaboration between ADI and the University of Leeds, funded by Innovate UK (project reference number 133522). It was agreed that the author would support this project to provide a practical study to supplement this research, while also helping ADI benefit from academic research relevant to their product development. The majority of the work involved redesigning the software in order to improve efficiencies for ADI's installation of the MyPathway application in the MSK department of STHT. Working as part of the team provided the author with the opportunity to explore some of the issues and challenges associated with working with raw, uncurated healthcare data. It also presented the opportunity to identify some important design considerations for process-aware healthcare systems.

During this study, the data is stratified by patient health outcome which is determined by comparing patient-reported outcome data over time. Process mining techniques were applied to the different event logs to identify potential indicators of a good or bad health outcome.

The work presented in this chapter adds strength to the thesis by building on the knowledge gained in Chapter 5, with a completely different set of opportunities and challenges to help answer the main research question posed in Chapter 1.

6.2 Stage 1: Planning

During the planning stage, a relationship was established with ADI. As part of this agreement, provision was made for the author to use the anonymised patient data from the MyPathway application, along with business and technical resources for research purposes within this study. In addition to these resources, two external programmers assisted with writing of the data extraction and transformation code. Ethical approval was granted (MREC17-108) via the Medicine and Health University Ethics Review team on 14/02/2020.

There were two aims to this study. The first was to determine how process mining techniques could be applied to the MyPathway data in order to identify possible indicators of a patient health outcome using PROM data. The second aim was to identify where changes could be made to the MyPathway system in order to improve data collection for future process mining studies. Towards these aims, six study-specific research questions were composed, these were:

1. Is it possible to determine measurable health outcomes for knee and spinal pain patients within the MyPathway data?

2. Can process mining techniques be applied to data from knee and spinal pain patients using the MyPathway application?

3. Do indicators of a health outcome exist in the MyPathway data for patients diagnosed with knee and spinal pain and if so can they be identified using process mining techniques?

4. Do demographic factors such as age or sex correlate with good or bad health outcomes?

5. What features, if any, of the care pathway correlate with good or bad health outcomes?

6. What can be changed in the MyPathway system in order to improve data collection for future process mining studies?

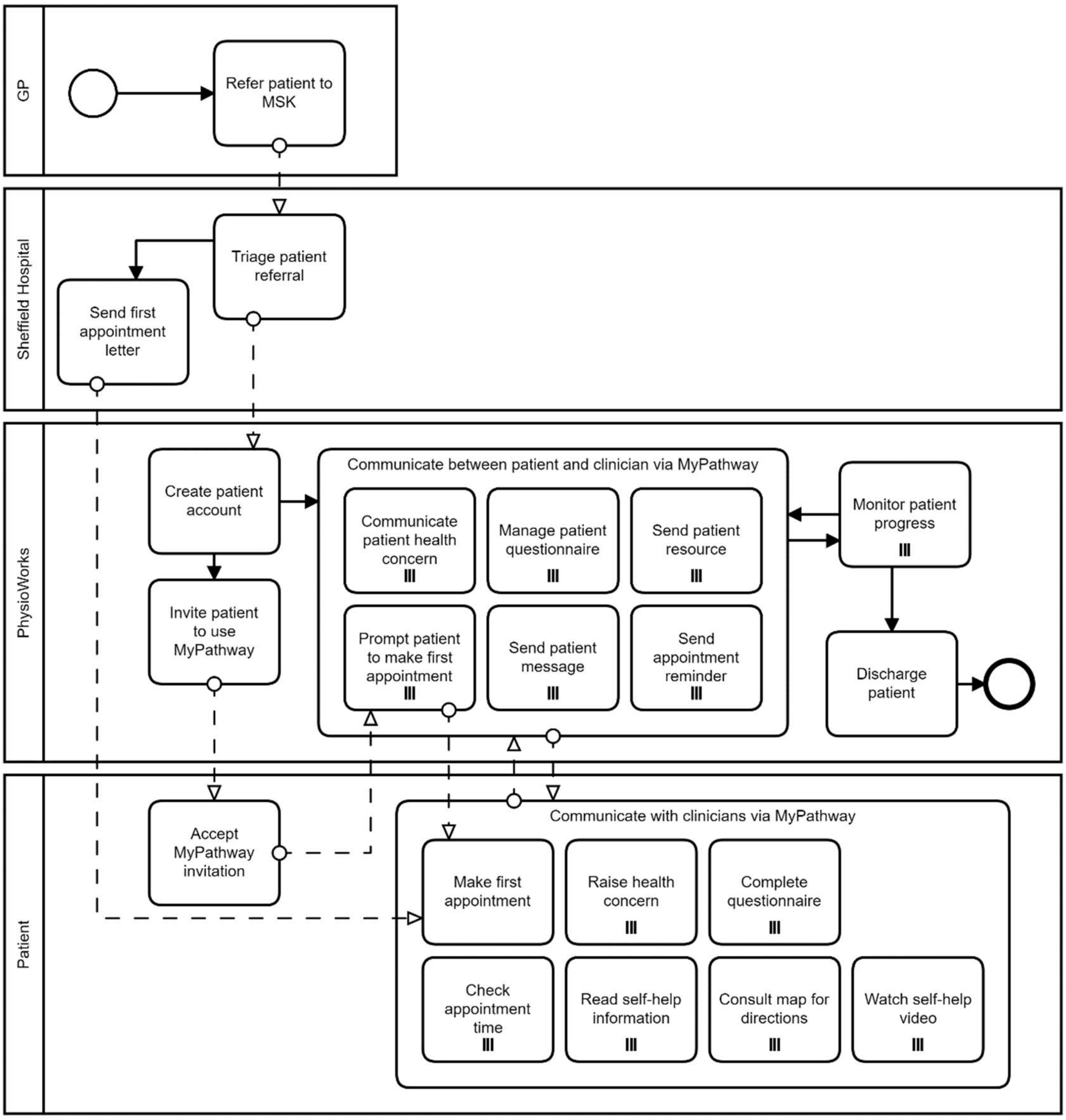
6.3 Stage 2: Extract, Transform and Load

Activities during this stage involved the gathering of detailed process and data related information. This included process information from the MyPathway application (see Section 6.3.1) and from the MSK department at STHT (see Section 6.3.2). The type and structure of the data was investigated to specify the requirements for the data extract. A health outcome was defined based on information from the EQ-5D questionnaire over time (see Section 6.3.3). Transformation of the data for this study was separated into two parts, data cleansing and data processing.

6.3.1 Overview of the MyPathway system

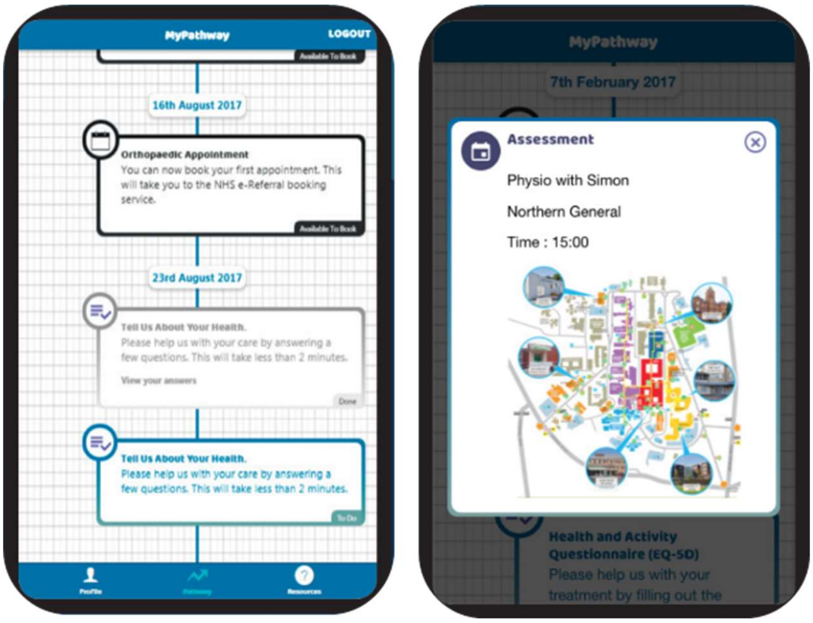
The MyPathway system is an example of a successful UK healthcare application and was developed by ADI in 2016. MyPathway connects patients to secondary care providers working at a hospital or a community based service such as PhysioWorks at STHT. The application has operational deployments in MSK services, mental health, chronic pain and motor neurone disease. The largest patient base is in the MSK department of STHT, where over 47,000 patients are referred each year. The BPMN diagram displayed in Figure 6.1 presents a high-level overview of the MyPathway process as implemented within the MSK department of STHT. The model was produced following a number of discussions with the ADI staff and after studying the existing process models. The process begins with a GP patient referral and ends when the patient is discharged.

Figure 6.1 High level process diagram of the MyPathway care process


GP
Refer patient to MSK
Sheffield Hospital
Triage patient referral
Send first appointment letter
PhysioWorks
Create patient account
Invite patient to use MyPathway
Communicate between patient and clinician via MyPathway
Communicate patient health concern
Manage patient questionnaire
Send patient resource
Prompt patient to make first appointment
Send patient message
Send appointment reminder
Monitor patient progress
Discharge patient
Patient
Accept MyPathway invitation
Communicate with clinicians via MyPathway
Make first appointment
Raise health concern
Complete questionnaire
Check appointment time
Read self-help information
Consult map for directions
Watch self-help video

The MyPathway application provides functionality that enables patients to communicate with clinicians regarding a health concern, check appointment times, complete questionnaires, watch self-help videos, refer to maps for appointment details and to read useful information related to their condition. Two MyPathway screen shots are presented in Figure 6.2. The first displays a patient's care pathway using a timeline of events, and the second displays patient appointment details [34].

Figure 6.2 MyPathway user interface displaying the patient's timeline and appointment details (reproduced with permission from ADI)



16th August 2017

Orthopaedic Appointment
You can now book your first appointment. This will take you to the NHS e-Referral booking service.

23rd August 2017

Tell Us About Your Health.
Please help us with your care by answering a few questions. This will take less than 2 minutes.
View your answers

Tell Us About Your Health.
Please help us with your care by answering a few questions. This will take less than 2 minutes.

7th February 2017

Assessment
Physio with Simon
Northern General
Time : 15:00

Health and Activity Questionnaire (EQ-5D)
Please help us with your treatment by filling out the

The application provides patients and their carers with 24 hour access to information related to their MSK condition. During an interview, the MSK Clinical Services Manager at STHT stated that the MyPathway application gave many patients peace of mind, as they felt in control and actively able to manage their condition [377]. Resources available via the application, such as questionnaires and self-help information can be tailored to a patient’s care pathway and automatically triggered by rules specified by the healthcare staff. The provision of PROMs in the form of questionnaires is an important feature of the application. Questionnaire responses from before and after a course of treatment can be compared to assess the efficacy of an intervention. In addition, the potential for analysis of disease specific PROMs issued during the course of a patient’s treatment, allows for early intervention and amendments to the treatment plan.

6.3.2 MSK care pathways at STHT

Information relating to MSK care pathways at STHT was gathered via a number of formal meetings. These meetings included members from the ADI development team, STHT MSK clinical and managerial staff. In addition to these meetings, four one-to-one interviews were carried out with the MSK Programme Manager and the PhysioWorks Clinical Service Manager at STHT (transcripts are available on request). The remainder of this section summarises the process discovered through these interviews.

Patients are referred to an MSK speciality from their GP. These specialities include Orthopaedics, Rheumatology, Pain, Physiotherapy and Therapy Services. Orthopaedics is a surgical speciality where surgical procedures are performed. Medication is prescribed along with therapies and life style changes in the Rheumatology speciality. The Pain clinic have a range of tools including facilities for guided injections, pain management programmes and the prescribing and de-prescribing of drugs that are outside the remit of a physiotherapist. Services provided by Physiotherapy, also referred to as PhysioWorks when located outside the hospital in the community, include the manipulation of joints, exercise programmes, administration of simple soft tissue injections and the prescription and de-prescription of a limited range of drugs. Care given by Therapy Services is similar to that given by PhysioWorks. The difference being, Therapy Services is located within a hospital and often has its own physiotherapy department where patients may attend post consultation or surgery. Most MSK specialties reside within the hospital. In the majority of cases, patients are referred directly to the PhysioWorks service. Approximately 88 percent of these patients remain in PhysioWorks for the duration of their treatment, though approximately 12 percent are re-referred into secondary care [377]. Due to the amount of data available, this study analyses data from patients attending the PhysioWorks service.

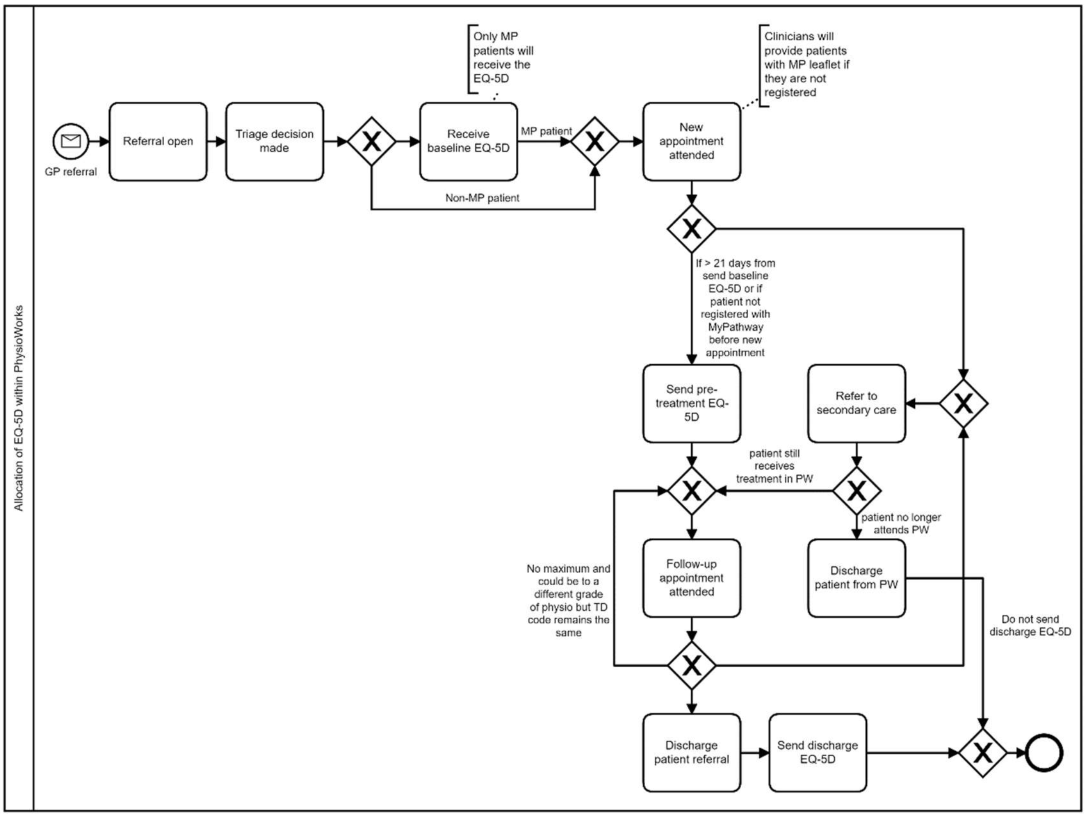
Physiotherapists at STHT believe that early intervention, management of patient expectations and the empowerment of patients, in order for self-management, are key factors for a good health outcome. Within PhysioWorks, patient care is approached in two ways: 1) by addressing the physical elements of the condition to improve the pathology and 2) by increasing the patient's level of activation, in terms of empowering patients to carry out activities that will enable repair. Patients that require high activation levels are often those where the pathology cannot repair, requiring them to self-manage their condition. Patients suffering regular flare-ups are also likely to require high activation levels to help prevent them from re-entering the system on a frequent basis.

Patients are triaged dependant on their biomechanical and biopsychosocial problems. During the triage process patients are referred to a clinic within a speciality. Biopsychosocial problems take into consideration the patient's social circumstances, for example, the patient's support mechanism, or their ability to work. The psychological element considers whether the patient's condition is significantly impacting on their mental health, or vice-versa. Patients are allocated a pathway via a clinic code. All clinic codes include the grade of the clinician and the part of the body for which the patient was referred, for example 'Clinic PW5 Knee' refers to PhysioWorks, physiotherapist grade 5, knee. The grading system for a physiotherapist begins at level five, progressing to level seven, Enhanced Role Physiotherapist (ERP), also known as Advanced Practice Physiotherapist (APP), Extended Scope Physiotherapist (ESP), then finally Integrated Pain Team (IPT). Patients are managed by the ERP and IPT clinicians when the biological part of their condition is not dominant. Different waiting times are associated with the different physiotherapist grades. Often patients triaged to an ERP or an IPT have a much shorter wait, than those referred to a more junior grade.

6.3.3.1 The EQ-5D process within MSK at STHT

Patient reported outcome measures, such as those collected via the EQ-5D, contribute to the information physiotherapists consider when assessing patients. Pathology specific and general questionnaires are allocated electronically through the MyPathway application. Pathology specific questionnaires such as the Keele STarT Back Screening Tool (SBST) [378] are used to measure the risk of chronicity and the Oswestry Low Back Pain Disability questionnaire [379] is used to measure the impact of the pain. Data obtained from these questionnaires is used to help decide on the grade of physiotherapist and package of care required by each patient. The generic EQ-5D questionnaire (Section 2.3.2.1) is issued to all patients via the MyPathway application regardless of their condition. If upon analysis of this data a patient's pathology appears to remain static, this could be due to the second questionnaire been issued too early or because no further improvements can be made. EQ-5D questionnaires are allocated at specific times during a patient's care pathway and are labelled accordingly. The BPMN diagram in Figure 6.3 has been created to demonstrate this process.

Figure 6.3 Allocation of the EQ-5D within PhysioWorks at STHT


Allocation of EQ-5D within PhysioWorks
GP referral
Referral open
Triage decision made
Receive baseline EQ-5D
Only MP patients will receive the EQ-5D
MP patient
Non-MP patient
New appointment attended
Clinicians will provide patients with MP leaflet if they are not registered
If > 21 days from send baseline EQ-5D or if patient not registered with MyPathway before new appointment
Send pre-treatment EQ-5D
Refer to secondary care
patient still receives treatment in PW
patient no longer attends PW
No maximum and could be to a different grade of physio but TD code remains the same
Follow-up appointment attended
Discharge patient from PW
Do not send discharge EQ-5D
Discharge patient referral
Send discharge EQ-5D

The process begins with the receipt of a GP patient referral before the patient is triaged into the PhysioWorks speciality. Any patient registered to use the MyPathway application receives a baseline EQ-5D questionnaire before attending their first appointment. After this appointment the patient will either be referred into secondary care or they will remain within the PhysioWorks speciality. If the patient attends their first appointment more than 21 days after receiving their baseline EQ-5D they are sent a pre-treatment EQ-5D. Patients may be discharged after their first appointment. However, this scenario is not modelled as these patients do not receive a second EQ-5D and therefore no health outcome can be determined. Patients referred into secondary care, not continuing in PhysioWorks will be discharged from the service.

Those patients do not receive a discharge EQ-5D. PhysioWorks patients may attend a number of follow-up (subsequent) appointments until they are either discharged or referred out of the service. Those discharged from the service will receive a discharge EQ-5D. There are exceptions to this, though these are not included as occurrences are rare.

6.3.3 Defining a health outcome using Patient Reported Outcome Measures

The EQ-5D is a well-tested questionnaire for the measurement of health (see Section 2.3.2.1) and was available in large volumes within the MyPathway system. Analysing data from the descriptive section of the questionnaire, in the form of health profiles is the recommended method for non-financial analysis [114]. Applying the Paretian Classification of Health Change (see Section 2.3.2.2) method to health profile data provides a simple and effective way to determine patient health change [119]. In order to explore possible indicators of health outcomes using the MyPathway data, the PCHC method was applied to health profile data from the EQ-5D questionnaire. Patients were identified and separated into four datasets dependant on this result.

When defining a health outcome during this study, the two health states from a matched pair of patient referral EQ-5Ds were compared. The first questionnaire must have been of type ‘baseline’ or ‘pre-treatment’, in other words before the start of treatment. The second questionnaire must have been of type ‘EQ5D Q’, ‘physioDischarge’ or ‘discharge’ and have occurred at least 22 days after the first appointment, to allow for a minimum of three weeks’ worth of treatment. The following logic was applied to the data:

- A health outcome is classified as ‘**Declined**’ if at least one of the five elements of the health state on the second EQ-5D is more than on the first EQ-5D and none are less.

- A health outcome is classified as ‘**No change**’ if all five elements of the health state for the first EQ-5D are equal to all five elements of the health state for the second EQ-5D.

- A health outcome is classified as '**Mixed**' if at least one of the five elements of the health state on the second EQ-5D is less than on the first EQ-5D and at least one is more.

- A health outcome is classified as '**Improved**' if at least one of the five elements of the health state on the second EQ-5D is less than on the first EQ-5D and none are more.

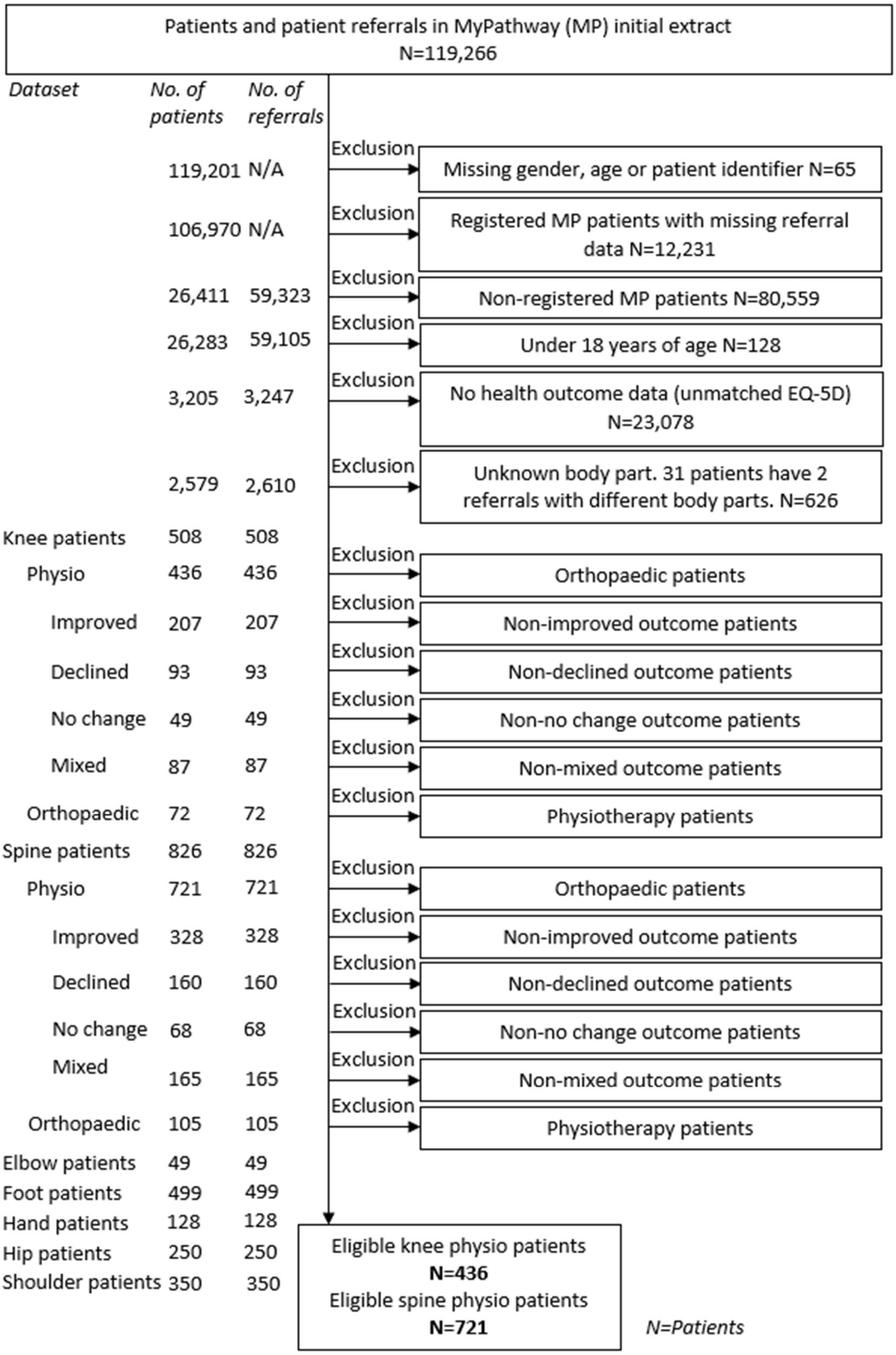
6.3.4 Data extraction

During the data extraction process, business knowledge was gained by collecting and understanding process models for the MyPathway application. Existing process models were obtained for many of the business scenarios, such as on-boarding (registration of patients to the MyPathway application). After discussions with the business experts, the author created models for processes that were either not available or available but with insufficient detail. The MyPathway data was examined, noting how and what type of information was stored (Section 4.2.2). Candidate timestamped events were identified and discussed with the research team. All data was stored in a NoSQL, MongoDB database (Section 3.1.2) with no schema. A logical data model was created by the author, specifying all data items and linkage requirements for the data extract (Section 4.2.2). Due to the non-standard nature of the query language, only one ADI developer was available on a part-time basis to perform the data extraction.

After delivery of the first data extract it was evident that many data quality issues existed. The main issue being the inability to trace a patient referral from beginning to end. Due to this problem the author adopted a lead role in the data extraction process. Twenty-five iterations were performed. Each iteration began with the delivery of a new data extract file, followed by extensive testing. Testing revealed a range of data issues corresponding to different patient scenarios. These scenarios were often complex and needed to be fully understood to design work-around solutions. Walk-through sessions were chaired by the author in order to evaluate these solutions and included technical and business experts from ADI and clinical and managerial staff from STHT. Following these sessions, new data extract requests were submitted to ADI. Due to limited resources, pressures from operational support and from new developments, many parts of these requests were unfulfilled. Therefore, in addition to the changes to the data extraction code, additional data items required for inference purposes during the data cleansing process were requested.

Typically, for a data extract of this size, when stored in a relational database, the extraction process should take no more than a week. This process took place over an eight month period. A breakdown of the issues is available in Appendix C. Figure 6.4 presents the number of patients included in the final extract, along with patient numbers and associated patient referrals at each stage of exclusion within the data transformation stage.

Figure 6.4 STROBE diagram showing number of patients and referrals


Patients and patient referrals in MyPathway (MP) initial extract
N=119,266
Exclusion
Missing gender, age or patient identifier N=65
Registered MP patients with missing referral data N=12,231
Non-registered MP patients N=80,559
Under 18 years of age N=128
No health outcome data (unmatched EQ-5D) N=23,078
Unknown body part. 31 patients have 2 referrals with different body parts. N=626
Orthopaedic patients
Non-improved outcome patients
Non-declined outcome patients
Non-no change outcome patients
Non-mixed outcome patients
Physiotherapy patients
Orthopaedic patients
Non-improved outcome patients
Non-declined outcome patients
Non-no change outcome patients
Non-mixed outcome patients
Physiotherapy patients
Eligible knee physio patients
N=436
Eligible spine physio patients
N=721
N=Patients

Dataset	*No. of patients*	*No. of referrals*
	119,201	N/A
	106,970	N/A
	26,411	59,323
	26,283	59,105
	3,205	3,247
	2,579	2,610
Knee patients	508	508
Physio	436	436
Improved	207	207
Declined	93	93
No change	49	49
Mixed	87	87
Orthopaedic	72	72
Spine patients	826	826
Physio	721	721
Improved	328	328
Declined	160	160
No change	68	68
Mixed	165	165
Orthopaedic	105	105
Elbow patients	49	49
Foot patients	499	499
Hand patients	128	128
Hip patients	250	250
Shoulder patients	350	350

The original MyPathway data extract consisted of 119,266 patients. Any patients with missing gender, age, identifier or referral information were excluded. Patients not registered to the application, under the age of 18 or without two matching EQ-5D questionnaires were also excluded. If an associated part of the body could not be determined, the patient was excluded. Where patients had more than one referral for the same part of the body, only the first referral was used.

Counts of referrals were taken for patients with knee, spine, elbow, foot/ankle, hand/wrist, hip or shoulder problems. Based on data volumes, only knee and spinal pain patients were selected. These patients were referred to either the orthopaedic or physiotherapy speciality. Again based on data volumes, only physiotherapy patients were selected. The number of physiotherapy knee and spinal pain patient referrals for each health outcome type is presented on the STROBE diagram. The total numbers for knee and spine are shown in the bottom box.

6.3.5 Data transformation and load

Due to the extensive number of data quality issues left outstanding in the final data extract, the data transformation stage was separated into two parts, data cleansing and data processing.

6.3.5.1 Data cleansing

The work-around solutions designed at the end of the data extraction stage were converted into a set of data cleansing rules and are available as pseudo code in Appendix D. Due to the unexpected extent and complexity of these rules, additional programming resource was required to implement these rules. The software was written using the Python programming language through the Jupyter Notebook and the Java programming language using the Eclipse IDE. It is important to state, that at the time of data extraction the MyPathway system was in its infancy and many of the features were work-in-progress. In addition, some of the data cleansing issues listed below were caused as a result of extracting the data from its native system.

The Mans Data Quality Matrix, explained in Section 4.1.4.1, was used to evaluate the data quality of the final data extract. Table 6.1 presents the results for the MyPathway data assessment.

Table 6.1 Data quality matrix for the MyPathway data extract

	Case	Event	Relationship	C_attribute	Position	Activity name	Timestamp	Resource	E_attribute
Missing data	L	H	H	L	N	N	N	N	H
Incorrect data	N	H	L	L	N	H	H	N	L
Imprecise data			N	N	N	N	H	N	N
Irrelevant data	H	H							

These data quality issues are described below:

- **Missing Cases, value L:** Patients with missing identifiers.

- **Missing Events, value H:** Questionnaire events with no matching referral.

- **Missing Relationship, value H:** Referrals could not be established for patients with missing referral identifiers. Inpatient waiting list appointment events were not linked to inpatient admitted events.

- **Missing Case Attributes, value L:** Events linked to patient referrals that occurred before MyPathway.

- **Missing Event Attributes, value H:** As EQ-5D questionnaire types were not labelled, an additional column was provided in the extract containing the trigger rule name for the questionnaire event. Not all events had a rule name. Similarly outpatient appointment events were not labelled as ‘first’ or ‘follow-up’. Allocated questionnaires were not labelled as ‘original’, ‘first’ or ‘second reminders’. The MyPathway system was not aware of patient diagnoses, therefore the part of the body listed within the clinic code was used as a proxy.

- **Incorrect Events, value H:** The MyPathway system became fully operational on the 15/05/2017, therefore any data before this date was unreliable. For inference purposes, data prior to this date was included in the extract. Some inpatients were admitted and discharged more than once with the same identifier. This was due to administration errors caused by poor design of the hospital systems. Duplicate questionnaires were assigned, caused by repeated data in the source systems.

- **Incorrect Relationship, value L:** Pre-admit numbers, used for inference of appointment events, incorrectly pointed to multiple referrals.

- **Incorrect Case Attributes, value L:** Invalid patient identifiers, caused by a technical issue when writing the data extract file.

- **Incorrect Activity Name, value H:** Telephone appointments labelled as attended or departed had the same meaning.

- **Incorrect Timestamps, value H:** Inpatient discharge events contained admission time. Outpatient and telephone appointment events were recorded retrospectively. All timestamps were transferred in the wrong format due to a coding issue. Outpatient appointment tracking events had unreliable and meaningless timestamps.

- **Incorrect Event Attributes, value L:** Patient identifiers were corrupted when formatting the data extract file.

- **Imprecise Timestamps, value H:** Timestamps for events generated by batch programs were not in the correct chronological order.

- **Irrelevant Cases, value H:** Non-MyPathway and test patients existed in the data extract.

- **Irrelevant Events, value H:** Spurious events existed due to manual intervention in the MyPathway system by administrative and clinical staff. Meaningless triage events generated by a batch program. Events that existed pre-MyPathway had been migrated, though did not relate to a referral.

MyPathway welcome messages were sent to all patients upon registration to the MyPathway application.

6.3.5.2 Data processing and loading

In this study, the case notion was represented by a patient referral. The purpose of data processing is to produce event logs, using clean data, which are optimised for process mining. To maximise on the capabilities of process mining, once the data had been cleansed, the following data processing activities were performed:

1) Aggregating event data

Events were aggregated to eliminate unnecessary complexity within the process models. The aggregated event types are listed below.

- All map events included the URL for the Google Maps reference, resulting in over 100 different map types. All map types were named to 'map sent'.
- Outpatient appointment tracking events had unreliable and meaningless timestamps. Patient movements were unreliably tracked by administrative and clinical staff during their hospital visit. Therefore, all outpatient tracking events were merged and renamed to 'outpatient appointment attended'.
- 'outpatient appointment available to book' events are automatically triggered immediately after 'triage decision made' events, therefore these two events were merged.
- The 'phone appointment attended' and 'phone appointment departed' events have the same meaning, therefore were renamed to 'phone appointment'.

2) Enriching event data

To understand the control flow of the process, it was necessary to add a layer of detail and split the following event types:

- Outpatient appointment:
  - *new* (first appointment for a referral)

    - *follow-up* (subsequent appointments for a referral)

- MyPathway invitation:
    - *Original invitation*
    - *Invitation reminder*
- EQ-5D questionnaire:
    - *baseline EQ-5D* (triggered when the first appointment is available to book)
    - *pre-treatment EQ-5D* (triggered after the first appointment is attended)
    - *EQ-5D Q* (assigned after the start of treatment and before discharge)
    - *discharge EQ-5D* (triggered by a discharge event)
    - *physioDischarge EQ-5D* (triggered by a PhysioWorks discharge)
- Allocation of the five questionnaire types above:
    - *original <<EQ-5D type>>*
    - *first <<EQ-5D type>>*
    - *second reminder <<EQ-5D type>>*

3) *Excluding event data*

Only six event types were included within the event logs. These related to the following activities: 'referral open', 'new outpatient appointment', 'follow-up outpatient appointment', 'phone appointment', 'outpatient appointment did not attend' and 'outpatient appointment cancelled'. Any event types that were clearly not a predictor of a health outcome, for example the tracking of a patient through their hospital visit, an event that always happened, for example a welcome message, or an event that rarely happened, for example the cancellation of a telephone appointment were excluded from the event log. Excluded event types are listed below:

- EQ-5D questionnaire events, though necessary for the assignment of a health outcome to a patient, had no value to the process models;

- Any event type related to an inpatient

- Patient appointment booking events (no accurate timestamps as events were recorded retrospectively)

- ‘referral closed’ events occurred randomly after the first appointment was booked

- ‘referral discharged’ events were used inconsistently by the clinicians (these may refer to a patient discharge from secondary care or a patient transfer between specialties)

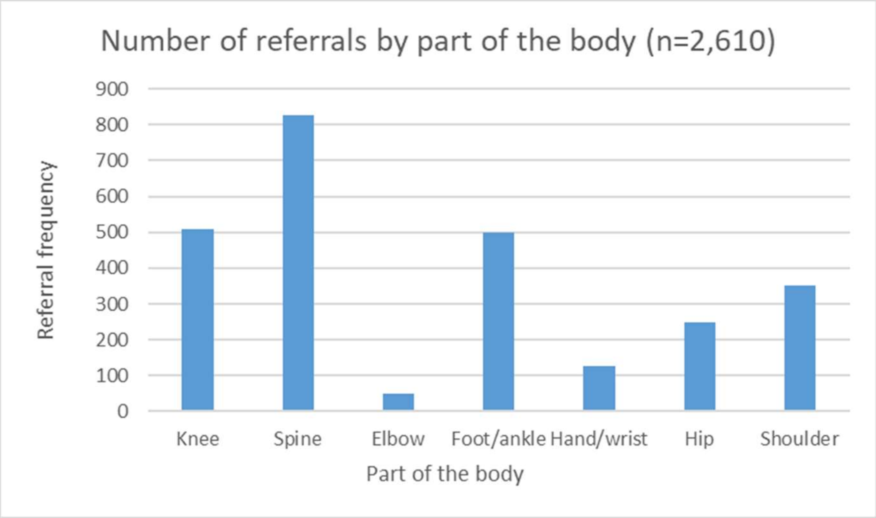
- The allocation of map events to patients always occurred before an appointment

The following six event types were selected for mining and analysis, these were: ‘Referral open’, ‘New outpatient appointment attended’, ‘Follow-up outpatient appointment attended’, ‘Outpatient appointment DNA’ (did not attend), ‘Outpatient appointment cancelled’ and ‘Phone appointment’. Any events after the date of the second matched questionnaire were removed, as these were not necessarily compliant with the allocated health outcome.

4) Filtering event data

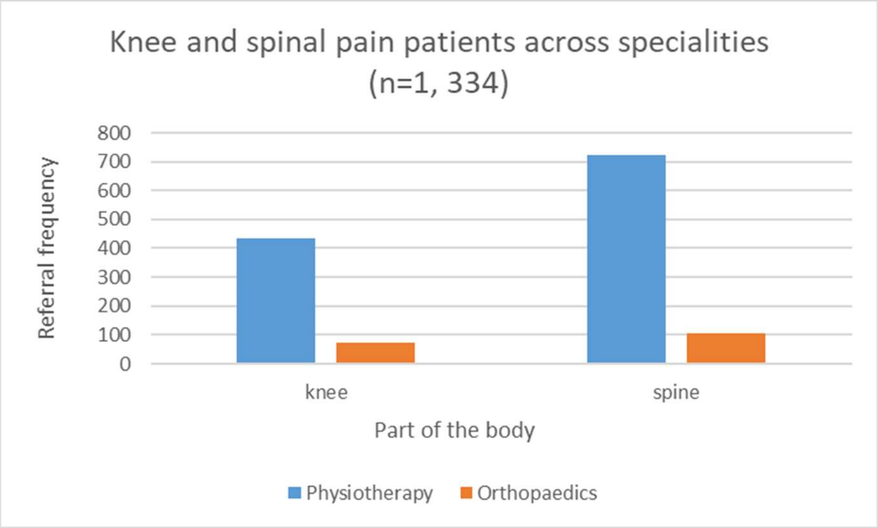
After the events had been aggregated, enriched and excluded, the event log contained 2,610 patient referrals. In order to perform effective process analysis, it was necessary to filter the data based on the patients’ treatment pathway. Figure 6.5 presents the number of matched EQ-5D patient referrals for different parts of the body.

Figure 6.5 Number of patient referrals for each body part



After considering the information in Figure 6.5 above, knee and spinal pain patients were selected for analysis, as they contained the highest number of matched patient referrals. The foot and ankle pain group of patients was not further investigated as it mainly consisted of podiatry patients which were out of scope for this study. The two chosen groups of patients were referred to either the physiotherapy or orthopaedics speciality. The number of patient referrals for each speciality can be seen in Figure 6.6.

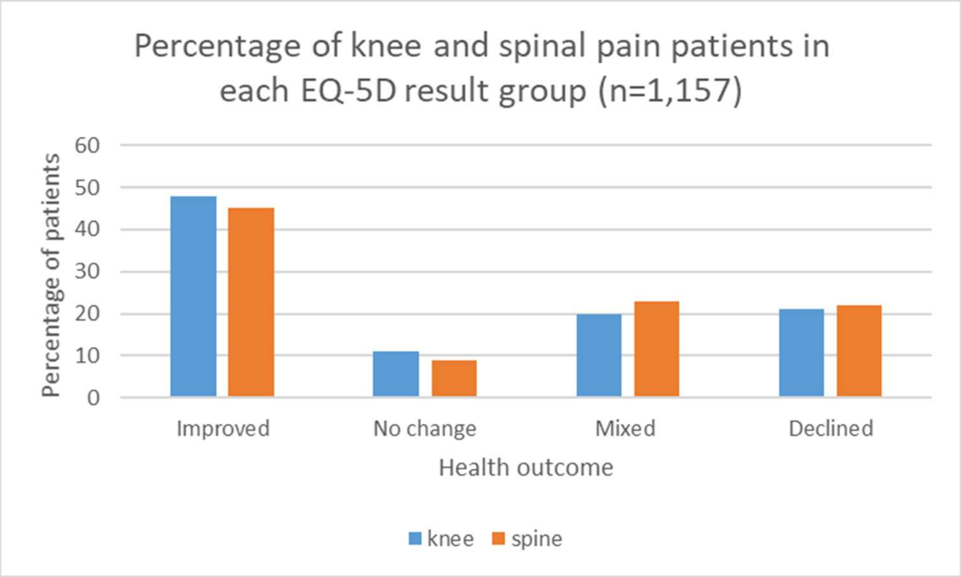
Figure 6.6 Number of knee and spinal pain patients by speciality



Due to the low number of patients referred to the orthopaedic speciality, only patients referred to the physiotherapy speciality were included in the analysis. Patients under the age of 18 years were excluded. Where patients had more than one referral for the same part of the body, only the first was used. Patients were then grouped by age group and sex. Due to the low number of patients, only two age groups were created, the under 50 years and 50 years and above. Age fifty was chosen, as it is a component of the American College of Rheumatology (ACR) criteria for the diagnosis of osteoarthritis of the knee [380]. In addition, fifty is close to the mean age of both knee (52 years) and spinal (50 years) pain patients in the study extract.

In total, 40 physiotherapy datasets were created for mining and analysis purposes. These datasets comprised of the following: one for each knee and spine health outcome (8); one for each knee and spine health outcome by sex and by health outcome (16); one for each knee and spine health outcome by age group and by health outcome (16). Two sets of files were created, the first for use within the R Studio and Microsoft Excel and the second for use with the process mining tool DISCO. The size of the files to be used within R and Excel were much smaller, as they did not contain any event details. The percentage of patients in each health outcome group is presented in Figure 6.7.

Figure 6.7 Breakdown of patients per health outcome group



As discussed in Section 2.3.2.2, it is important to check that the percentage of patients resulting in a ‘mixed’ health outcome is not dominant within the dataset. As is evident, this was not the case with this study, as 80 percent of the knee pain and 77 percent of spinal pain patients fell into the other three categories.

5. *Loading*

The functionality of both the Celonis and DISCO process mining tools are similar. In Chapter 5, DISCO was used to discover disease trajectory models using the MIMIC-III data, therefore, to enable the exploration of both tools, Celonis was chosen for the discovery of care pathways using the MyPathway data. The files were used in R and Excel to help characterise the datasets and the event logs were imported Celonis ready for process mining and analysis.

6.3.5.3 Redesign considerations for process mining studies

After reviewing the issues from the previous two sections, a list of redesign considerations in order to improve data collection for future process mining studies is presented below.

1. Ensure a current data model exists.

2. Ensure that all key data is captured by the system, such as patient diagnosis.
3. Generate an event log capturing all the necessary data as it is created in the system.
4. If necessary, enrich the data by creating new event types as they are entered into the event log.
5. Consider the technical architecture to ensure the performance of the operational system is not effected by the point 5 above.
6. Decide on the correct case classification, e.g., 'patient referral' to ensure all related items are linked, regardless of the underlying technology.
7. Record events in the event log at the correct level of abstraction.
8. Do not record data in the event log if the supporting business process cannot consistently capture the data accurately.
9. Ensure all case-related data is present when migrating historic data into a new system.

This list contributes to the considerations for healthcare system development and process mining research in Section 8.5.4.

6.4 Stage 3: Mining and analysis

When evaluating the results from this section, the PhysioWorks Clinical Service Manager at STHT requested that additional information on each dataset was provided in order to rule out any potential bias. This information is presented in Section 6.4.1 and includes patient sex and age related statistics, severity levels at baseline and the distribution of patients within clinics for all health outcome groups. Forty process models were discovered. The results from these models were analysed by comparing the data for patients with an improved and declined health outcome. Details for patients resulting in a 'no change' or 'mixed' health outcome were included for completeness purposes. The results from this analysis are presented in Section 6.4.2.

6.4.1 Data characterisation

An overview of the MyPathway dataset was presented in Section 1.4.2 and further described in Section 4.2.2. The following two sub-sections present data characteristics for the PhysioWorks knee and spinal pain patient datasets. Some of the reasons behind the provision of this information are discussed below.

The distribution of patient ages within a dataset can significantly influence the results. For example, sometimes, especially if the patient is frail and elderly, it may not be possible with physiotherapy to reduce their pain, however, it may be possible, if supplied with the correct mobility aids such as walking frames, shower rails and seats, to improve their activity level. A patients' age can also have an effect on appointment waiting times. Patients of retirement age are often more flexible with regard to appointment times, and can therefore accommodate appointments at short notice. As such, when appointment cancellations occur, these patients are often the first to be contacted by the PhysioWorks staff at STHT.

When analysing the results, it is important to be aware of the proportion of male and female, and young and old patients within each event log. In addition, the general health across each of the four patient health outcome groups should be known at baseline. The distribution of the five dimension scores (mobility, self-care, usual activities, pain/discomfort, anxiety/depression) is presented using boxplots. Note, that the standard EQ-5D scale runs between the severity levels one and five (Section 2.3.2.1), however within the MyPathway application the scales run between zero and four, with zero indicating no problems and four extreme problems. Data for patients with high scoring health profiles at baseline (indicating they were highly impacted) may explain shorter waiting times between GP referral and first appointment.

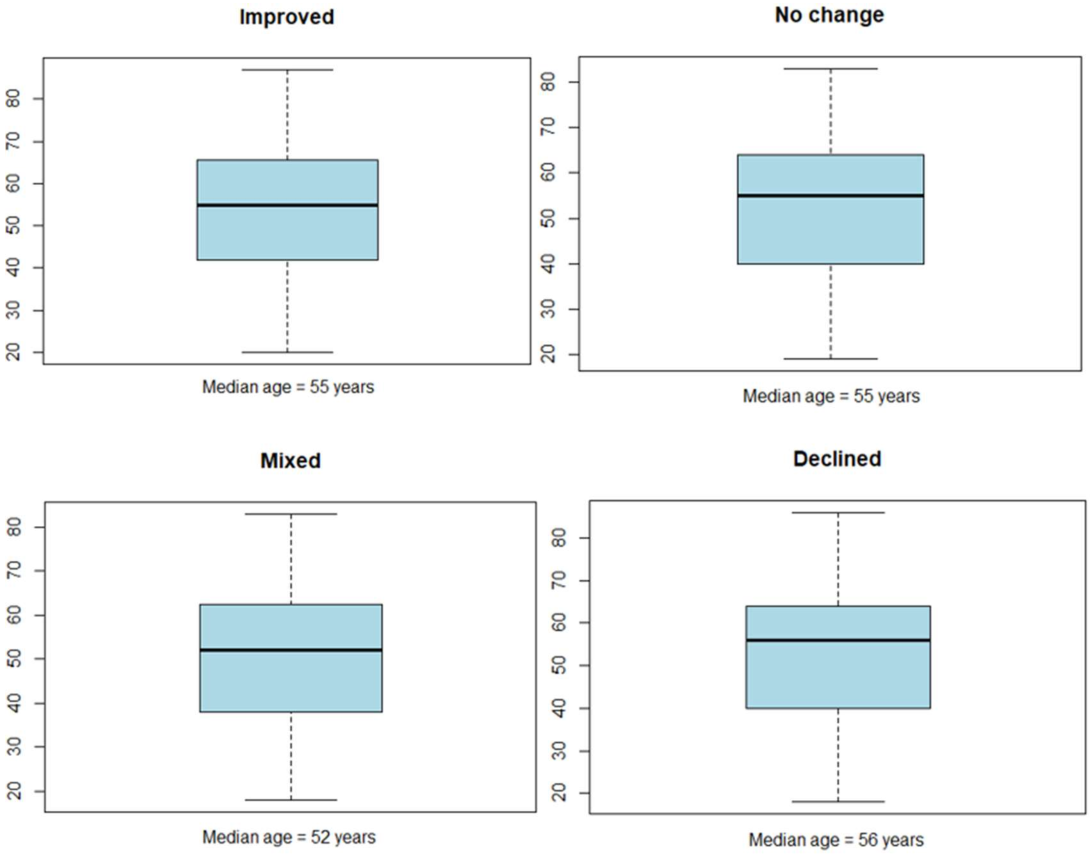
Within each health outcome group, consideration should also be given to the percentage of patients allocated to the different clinics. Within the PhysioWorks service at STHT, the clinic codes 'PW5' and 'PW6' are run by junior and lower grade physiotherapists. Patients receiving treatment in these clinics would not be offered an urgent appointment, unlike the clinics 'PW7 and 'PWERP', where urgent appointments are available. The PhysioWorks ERP clinics are run by highly specialised physiotherapists. Patients in these clinics often experience more complex problems. As ERPs often perform complex procedures, it is typical for patients to have telephone appointments in order for them to receive their test results.

The boxplots, charts and graphs in the following two sub-sections present the data characteristics deemed necessary to correctly interpret the knee and spinal pain patient results presented in Section 6.4.2.

6.4.1.1 Knee pain datasets

Figure 6.8 presents the age distribution for knee pain patients within all four health outcome groups.

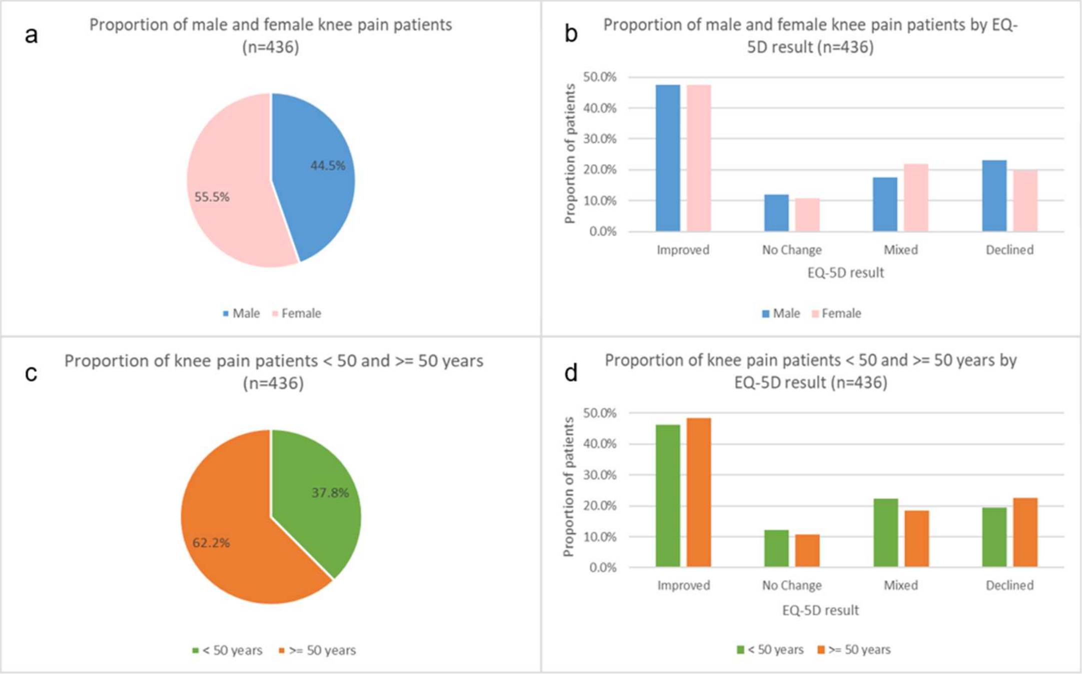
Figure 6.8 Age distribution for knee pain patients within the four health outcome groups



The information from these boxplots shows that no noticeable differences exist between the improved and declined datasets with regard to the minimum, median or maximum age of patients.

In Figure 6.9a, the percentage of knee pain patients by sex is presented. Figure 6.9b shows the same information, though divided by health outcome group. The percentage of knee pain patients by age group is presented in Figure 6.9c. Figure 6.9d shows the same information as Figure 6.9c, though divided by health outcome group.

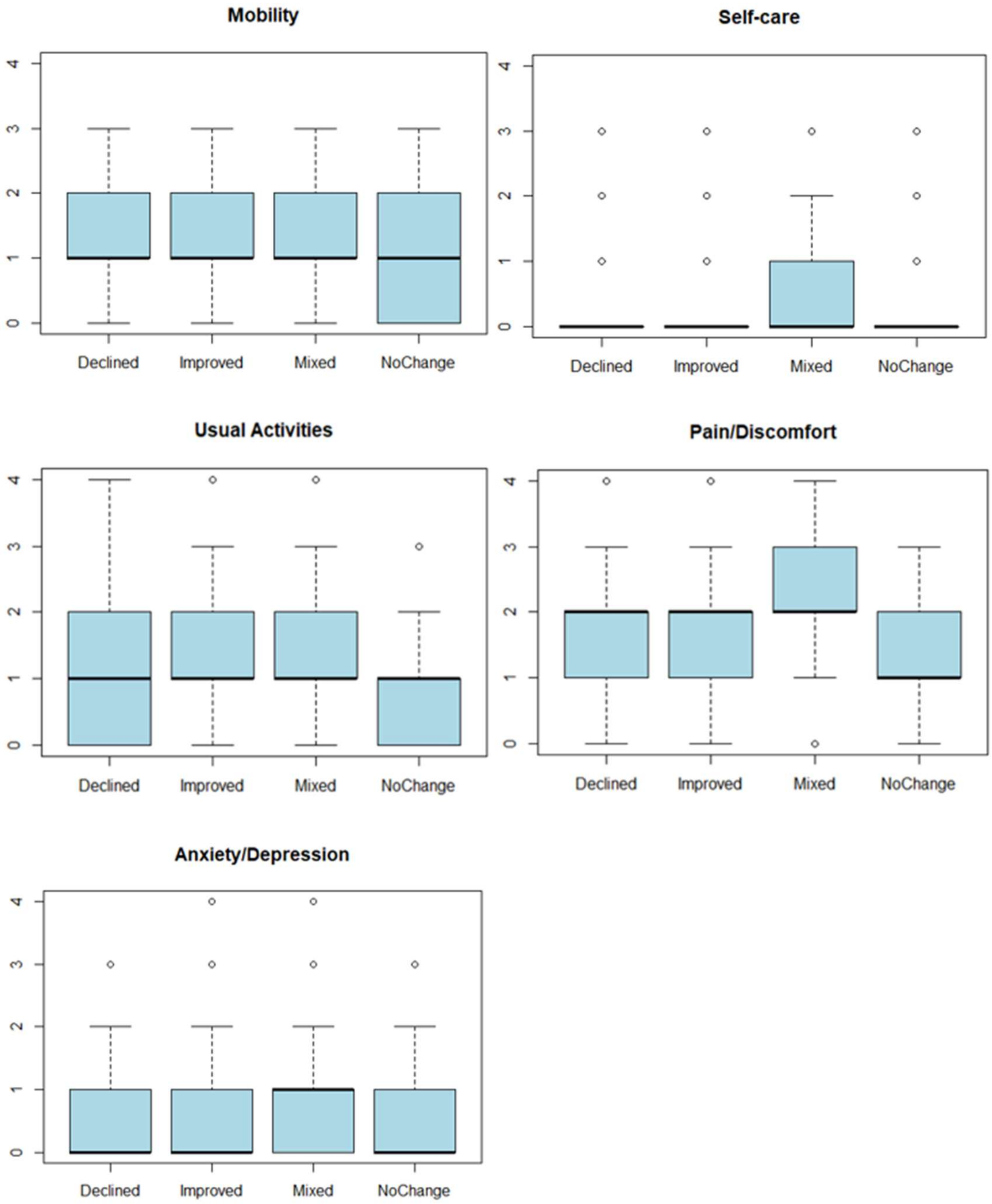
Figure 6.9(a) Knee pain patients by sex, (b) by sex and health outcome, (c) Knee pain patients by age group, (d) by age group and health outcome



The information presented in Figure 6.9a shows a slightly higher percentage of male patients (11%). Whereas the information in Figure 6.9c shows a higher percentage of patients in the older age group (24.4%). There are no noticeable differences when both datasets are further divided by health outcome.

Figure 6.10 presents the distribution of the severity levels for knee pain patients over each of the five EQ-5D dimensions at baseline.

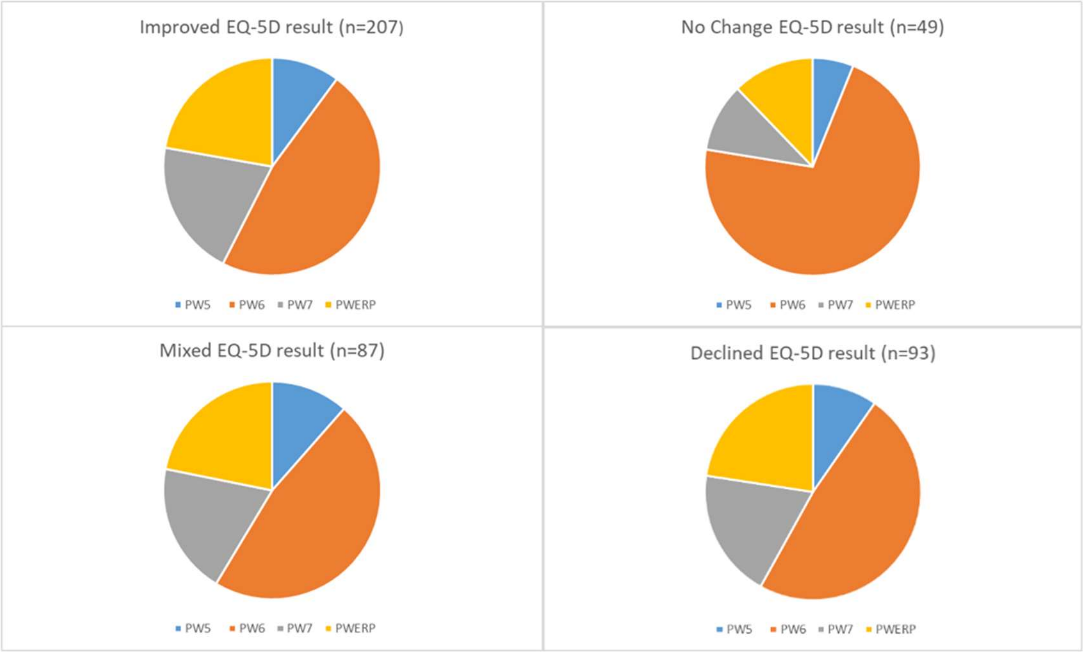
Figure 6.10 Distribution of severity levels for knee pain patients for each of the dimension scores at baseline by health outcome



As expected, for patients receiving physiotherapy treatment, the overall pain/discomfort and mobility dimensions score the highest. Unfortunately, unlike Kolotkin and Stratford (Section 2.3.2.1), it is impossible to predict whether the severity of the scores at baseline are likely to have influenced the patient health outcomes. This is because there is no noticeable difference for patients with an improved and declined health outcome.

The four charts in Figure 6.11 present information on the percentage of patients allocated to the four knee pain clinics within each health outcome group.

Figure 6.11 Distribution of clinics by health outcome

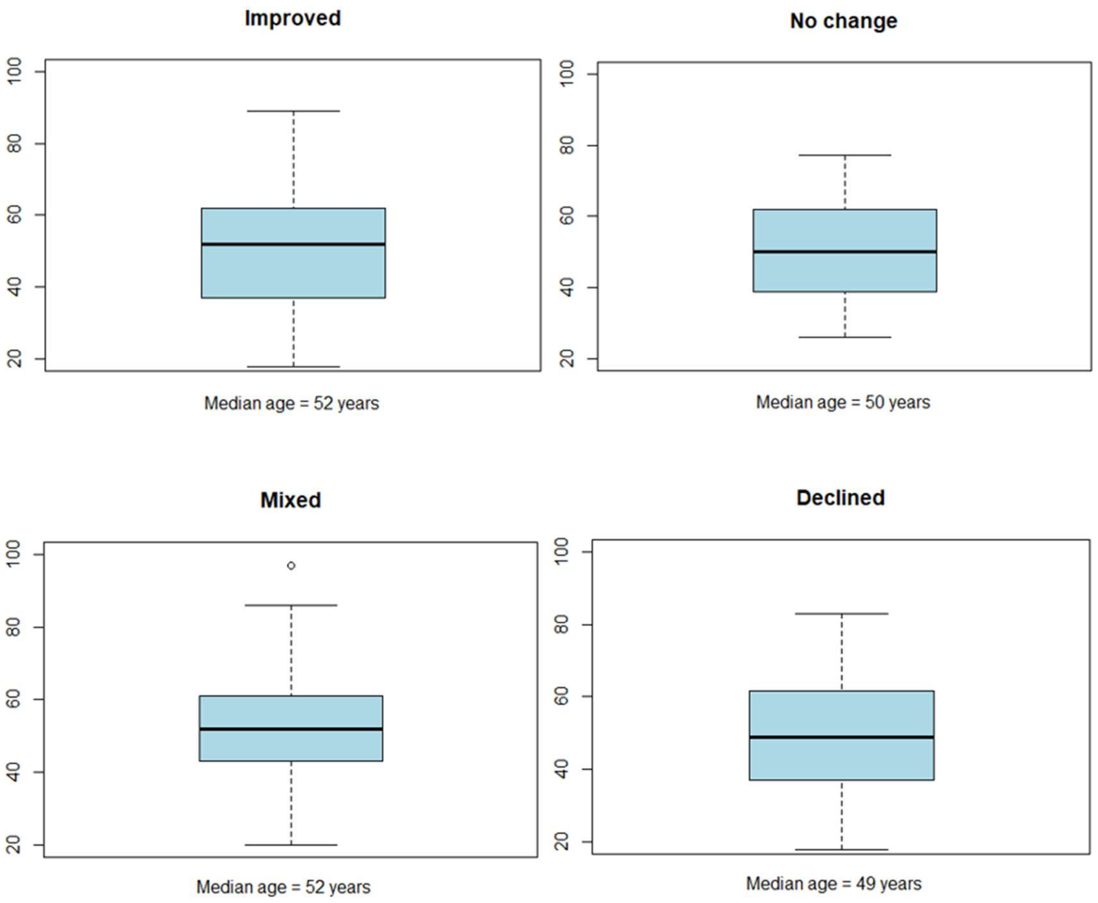


As is evident, the proportion of patients in the Improved and Declined health outcome groups are similar, only the percentage of patients in the no change group differs. When analysing the results in the following section, it must be remembered that patients attending this clinic would not be offered an urgent appointment.

6.4.1.2 Spinal pain dataset

Figure 6.12 presents the age distribution for spinal pain patients within all four health outcome groups.

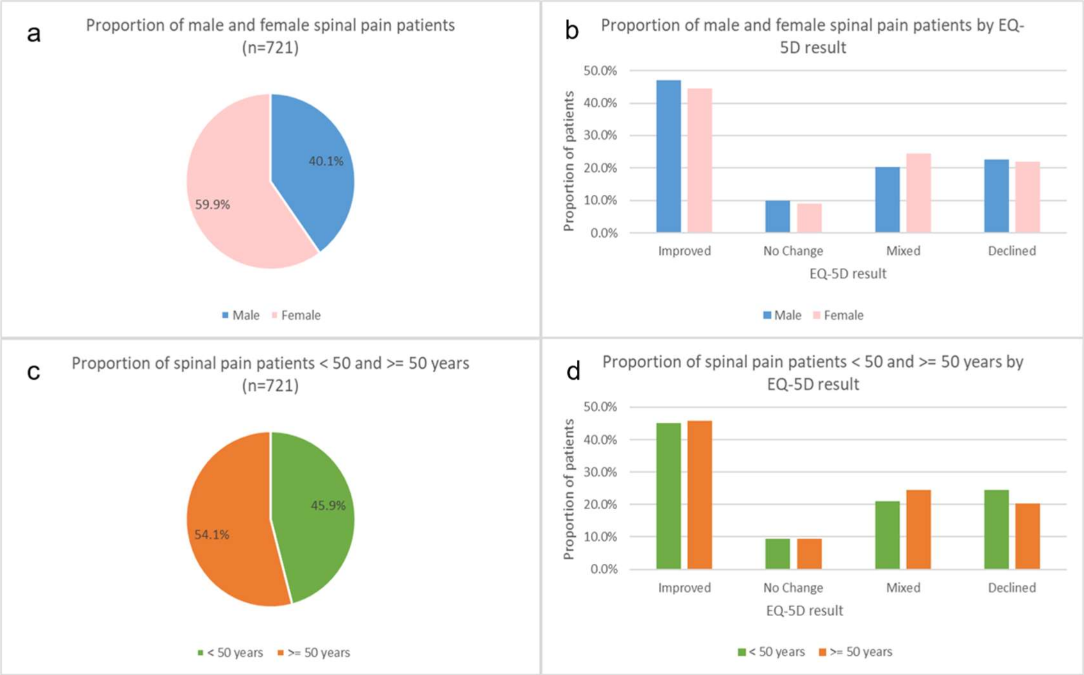
Figure 6.12 Age distribution for spinal pain patients within the four health outcome groups



The information from these boxplots show no noticeable differences exist between the improved and declined datasets with regard to the minimum and maximum age of patients. The median age of patients in the improved dataset is three years older than that of the declined dataset.

In Figure 6.13a the percentage of spinal pain patients by sex is presented. Figure 6.13b shows the same information, though divided by health outcome group. The percentage of spinal pain patients by age group is presented in Figure 6.13c. Figure 6.13d shows the same information as 6.13c, though divided by health outcome group.

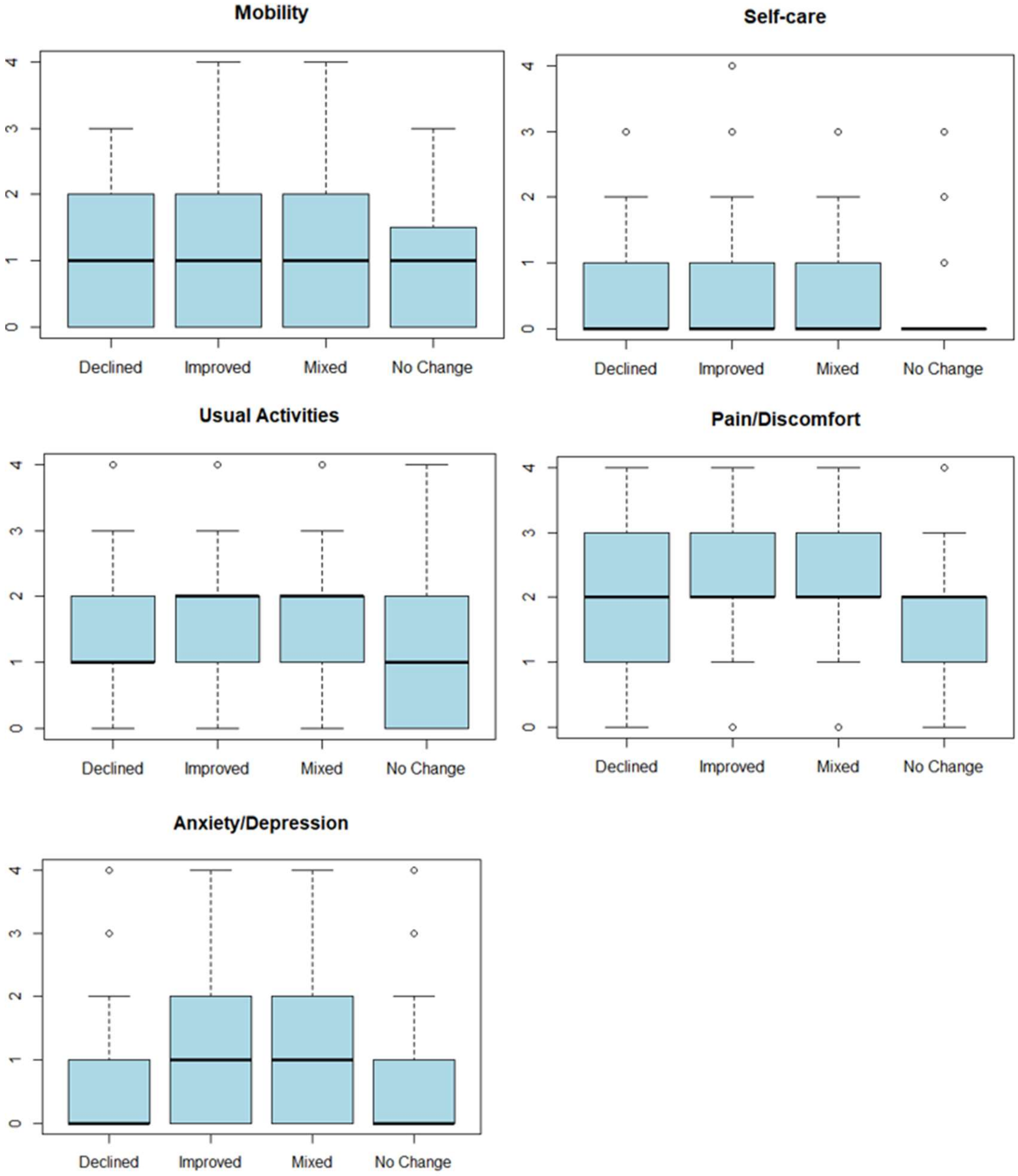
Figure 6.13(a) Spinal pain patients by sex, (b) by sex and health outcome, (c) Spinal pain patients by age group, (d) by age group and health outcome



The information presented in Figure 6.13a shows a higher percentage of female patients (19.8%). Whereas the information in Figure 6.13c shows a slightly higher percentage of patients in the older age group (8.2%). When the datasets are divided by health outcome, it is evident that a higher percentage of patients in the younger age group have a declined health outcome.

Figure 6.14 presents the distribution of the severity levels for spinal pain patients over each of the five EQ-5D dimensions at baseline.

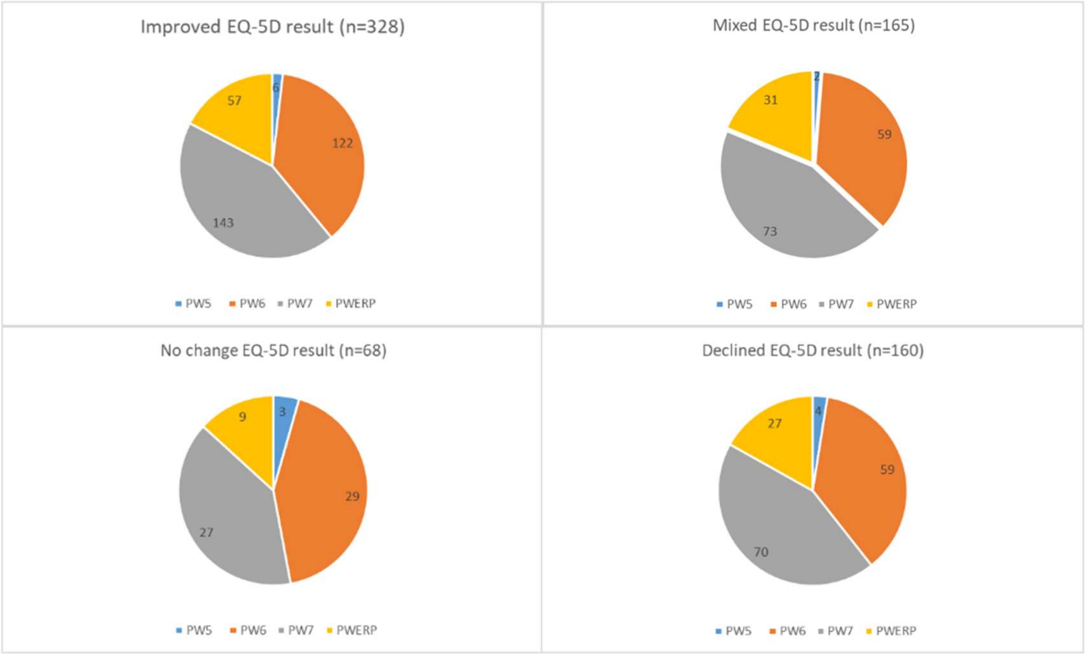
Figure 6.14 Distribution of severity levels for spinal pain patients for each of the dimension scores at baseline by health outcome



Overall, the severity levels for patients in the spinal pain dataset were higher than those in the knee pain dataset, with the exception of the mobility dimension, where spinal pain patients rated themselves as slightly more mobile at baseline. For spinal pain patients, the pain/discomfort dimension scored highest overall, where patients with an improved health outcome experienced slightly higher pain levels than those with a declined health outcome at baseline. This observation is similar within the dimension anxiety/depression.

The charts in Figure 6.15 present the percentage of patients allocated to the four spinal pain clinics for each health outcome group.

Figure 6.15 Distribution of clinics for spinal pain patients by health outcome



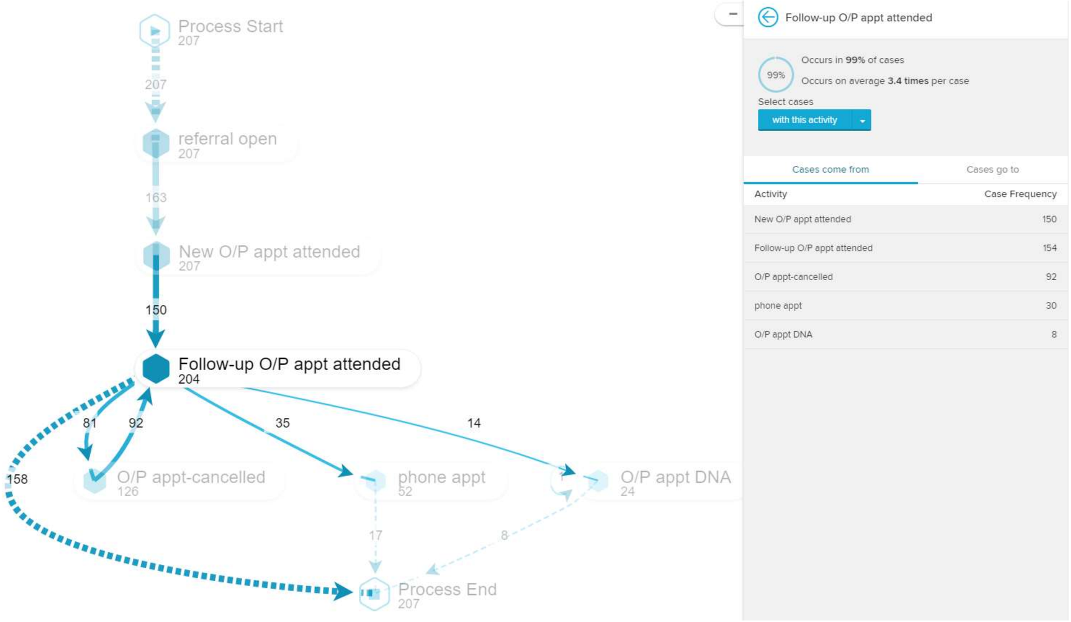
Similar to the patients in the knee pain dataset, the percentage of patients with improved and declined health outcomes are similar. Only the percentage of patients in the no change group differs. Again, it must be remembered that patients attending this clinic would not be offered an urgent appointment.

6.4.2 Process discovery and analysis

The exposures in this study are time interval between events and number of events per patient referral, age group and sex. Events refer to contact events, which include: follow-up outpatient appointments; telephone appointments; appointments where patients did not attend without cancelling; and outpatient appointment cancellations. The outcome variable is the health outcome, which is either improved, declined, no change or mixed. Potential confounders are identified after characterising the data in Section 6.4.1.

Forty process models showing case frequency and time related information were discovered for the PhysioWorks knee and spinal pain patients using the event logs created in Section 6.3.5. In order to report the results in a meaningful way, this data was consolidated and presented using a selection of tables and charts. The models were analysed to identify potential indicators of a health outcome. Indicators were found by comparing the percentage of contact events in the different event logs. In addition, the median number of times that each contact event occurred per patient was compared. An example of this is presented in Figure 6.16.

Figure 6.16 Process model showing that the Follow-up outpatient appointment attended event occurs in 99% of patients and for those patients it occurs on average (median) 3.4 times



Follow-up O/P appt attended

Occurs in **99%** of cases

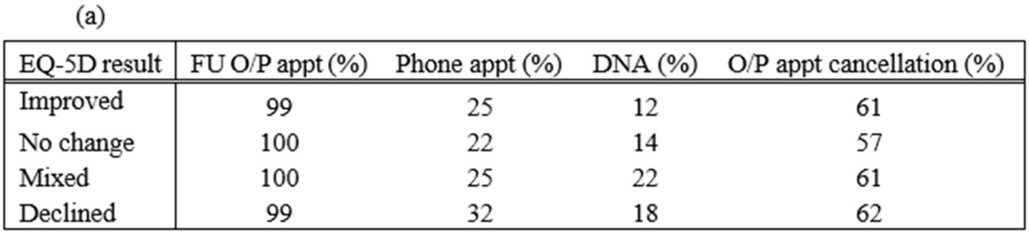
Occurs on average **3.4 times** per case

Select cases

Activity	Case Frequency
New O/P appt attended	150
Follow-up O/P appt attended	154
O/P appt-cancelled	92
phone appt	30
O/P appt DNA	8

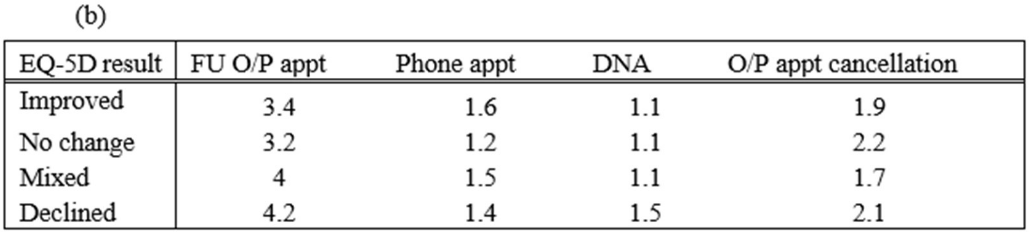
All process models showed a high degree of trace variation. On average, only 6.1 percent of all patient referrals followed the algorithmic typical path, which consisted of: *'referral open'*→*'New outpatient appointment attended'*→*'Follow-up outpatient appointment attended'*. This was not surprising, given the nature of the six event types included in the models. The models showed that patients would randomly cancel appointments, not attend appointments and attend telephone appointments throughout the core process. Information from these process models was recorded in tables, before transferring the results into bar charts for ease of reading. An example of how this data was recorded is presented using Table 6.2 and Figure 6.17 below.

Table 6.2(a) For all knee pain patients, percentage of times each contact event type occurs. (b) For all knee pain patients, median number of times each contact event type occurs per patient



(a)

EQ-5D result	FU O/P appt (%)	Phone appt (%)	DNA (%)	O/P appt cancellation (%)
Improved	99	25	12	61
No change	100	22	14	57
Mixed	100	25	22	61
Declined	99	32	18	62



(b)

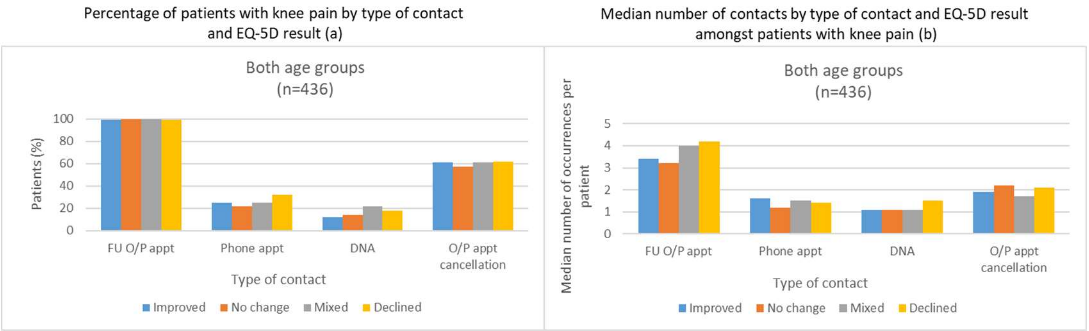
EQ-5D result	FU O/P appt	Phone appt	DNA	O/P appt cancellation
Improved	3.4	1.6	1.1	1.9
No change	3.2	1.2	1.1	2.2
Mixed	4	1.5	1.1	1.7
Declined	4.2	1.4	1.5	2.1

[1]

Table (a) displays the percentage of patients in the model with each type of contact event for each health outcome. Table (b) extends the information in table (a) to include the median number of times each contact event occurs per patient. After creating these tables for each model the information was transferred onto bar charts, as shown in Figure 6.17.

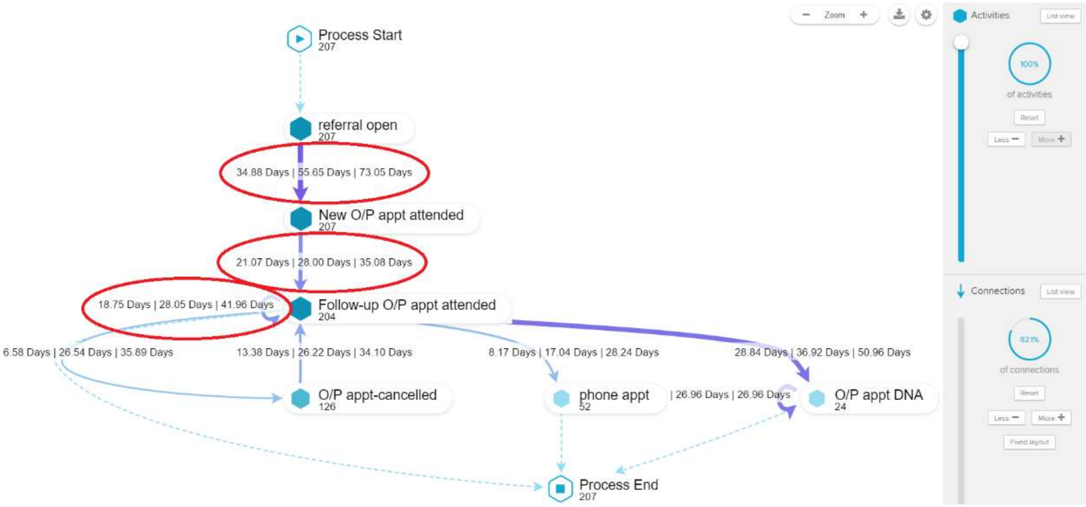
[1] The following abbreviations apply: FU = follow-up; O/P = outpatient; DNA = did not attend appointment

Figure 6.17(a) For all knee pain patients, percentage of times each contact event type occurs. (b) For all knee pain patients, number of times each contact event type occurs per patient



After comparing contact events, time related information was calculated and presented in tables. First, the overall case duration for each model was recorded using information from Custom KPI 1 (Section 4.1.3.1). Second, time intervals between contact events were compared, using Custom KPI 2 (Section 4.1.3.1). This process is demonstrated in Figure 6.18 and Table 6.3.

Figure 6.18 Process model showing the three time intervals between activities for knee pain patients with an improved EQ-5D result


Process Start
207
referral open
207
34.88 Days | 55.65 Days | 73.05 Days
New O/P appt attended
207
21.07 Days | 28.00 Days | 35.08 Days
18.75 Days | 28.05 Days | 41.96 Days
Follow-up O/P appt attended
204
6.58 Days | 26.54 Days | 35.89 Days
13.38 Days | 26.22 Days | 34.10 Days
8.17 Days | 17.04 Days | 28.24 Days
28.84 Days | 36.92 Days | 50.96 Days
O/P appt-cancelled
126
phone appt
52
| 26.96 Days | 26.96 Days
O/P appt DNA
24
Process End
207
Zoom
Activities
List view
100%
of activities
Reset
Less
More
Connections
List view
82.1%
of connections
Reset
Less
More
Fixed layout

The results were transferred into six tables for both the knee and spinal pain patients (one for each entire event log, which is presented twice to enable comparisons to be made, and two for each event log split by sex and age group). Table 6.3 presents an example of how this temporal information was transferred from the process models to the table for knee pain patients with an improved health outcome.

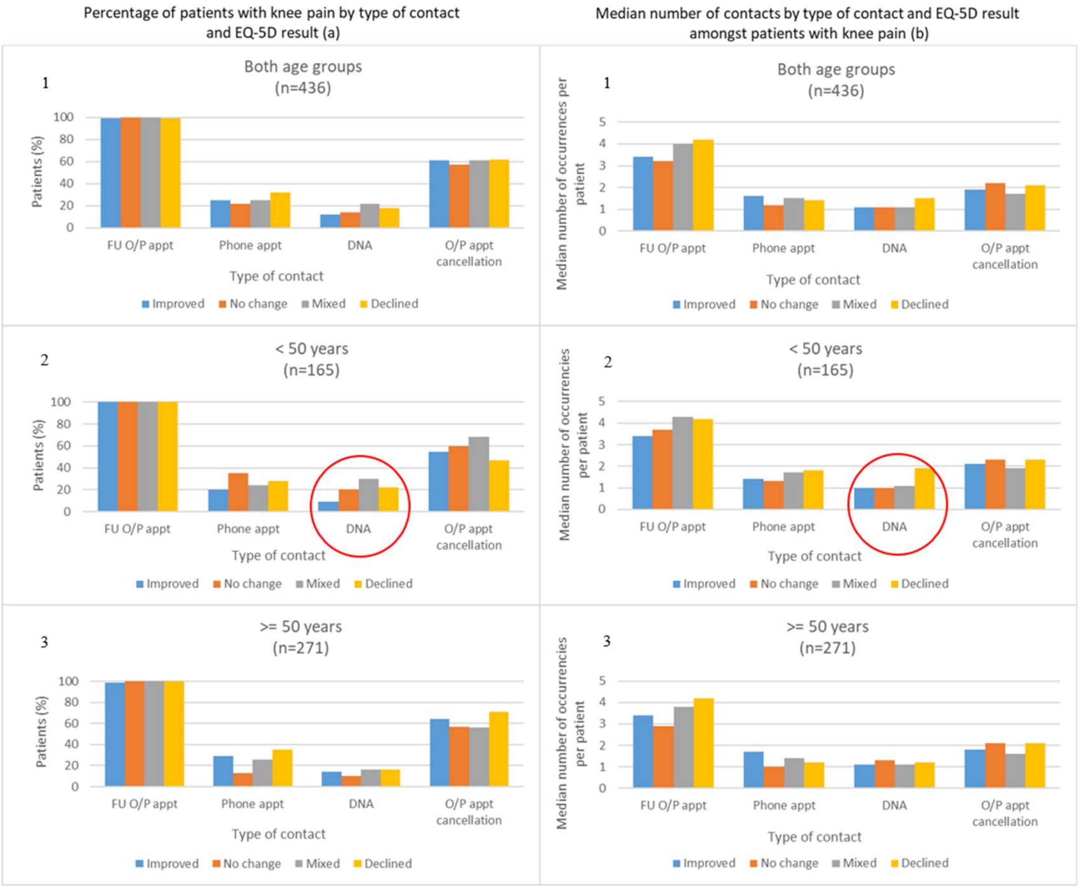
Table 6.3 Time information for knee pain patients with an improved EQ-5D result

EQ-5D result	Sample size (patients(%))	Case duration (days)			RO to 1st O/P appt (days(% of cases))			1st to 2nd O/P appt (days(% of cases))			O/P appt to O/P appt (days(% of cases))		
	n=436(100)	Q1	Median	Q3	Q1	Median	Q3	Q1	Median	Q3	Q1	Median	Q3
Improved	207(47)	118	177	251	34.88	55.65(79)	73.05	21.07	28.00(72)	35.08	18.75	28.05(74)	41.96
No change													
Mixed													
Declined													

6.4.2.1 Results for knee pain patients

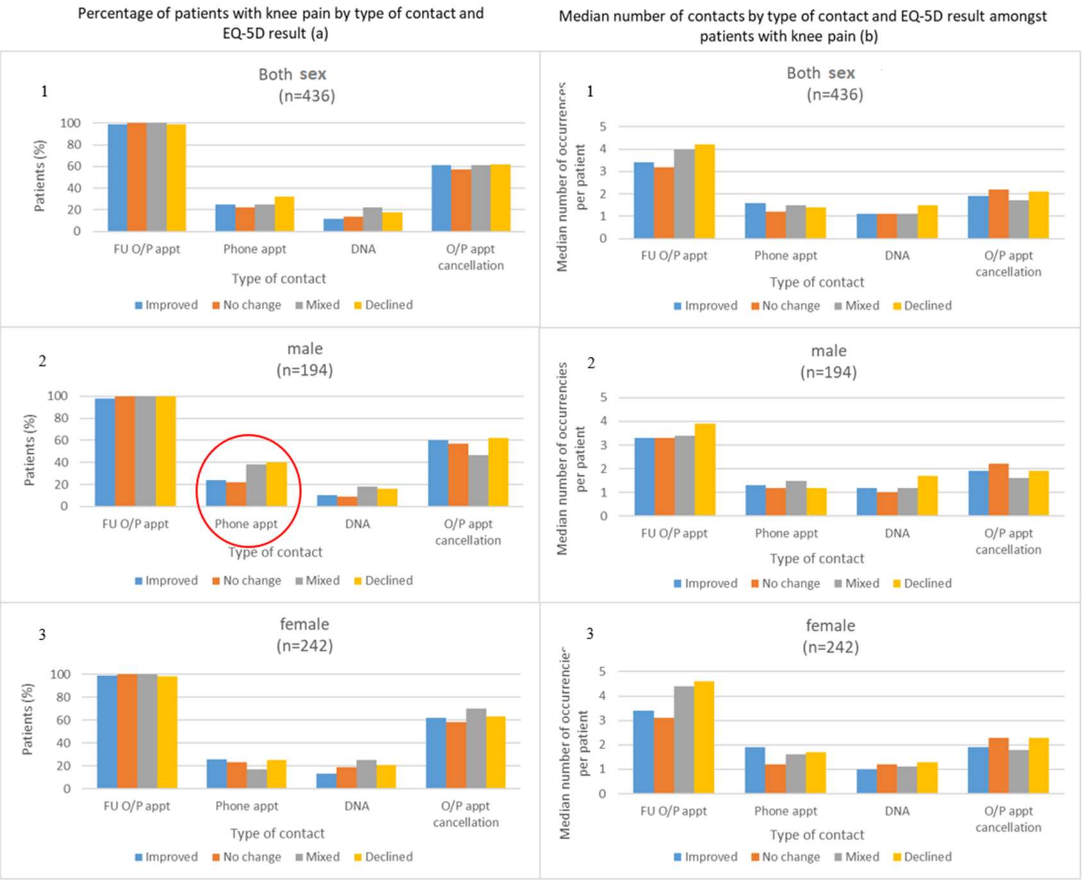
The information presented in Figures 6.19 and 6.20 relates to Figure 6.17. Figures 6.19/20(a) show the percentage of knee pain patients with each type of contact event for each health outcome. Charts 6.19/20(a)1 display the information for all patients and charts 6.19/20(a)2, display the information for patients in the younger age group and males, respectively. Charts 6.19/6.20(a)3 display the information for patients in the older age group and females, respectively. The charts in Figures 6.19/6.20(b) correspond to the information shown in (a) by presenting the median number of times each patient performs a specific content event.

Figure 6.19 (a) Proportion of knee pain patients with contact events by age group and health outcome, (b) for those patients, the median number of event occurrences



It is evident from Figure 6.19(b), that patients with a declined health outcome have, on average, more follow-up appointments than those with an improved outcome. Regarding the percentage of patients who did not attend their appointment, Figure 6.19(a)1 shows a six percent increase for patients with a declined health outcome compared to those with an improved health outcome. In addition, Figure 6.19(b)1 shows, that for those DNA patients with a declined health outcome, there is a higher number of occurrences per patient (0.4), compared to those that improved. When the event logs were divided by age group, the difference between the percentage of DNA patients with improved and declined health outcomes became more noticeable. In the younger age group (see Figure 6.19(a)2), there were over twice as many patients with a declined health outcome (22%) compared to those with an improved health outcome (9%). Figure 6.19(b)2 shows that, on average, patients with a declined health outcome missed almost twice as many appointments (1.9) as those that improved (1.0). Figure 6.20 presents the same information for when the dataset is split by sex.

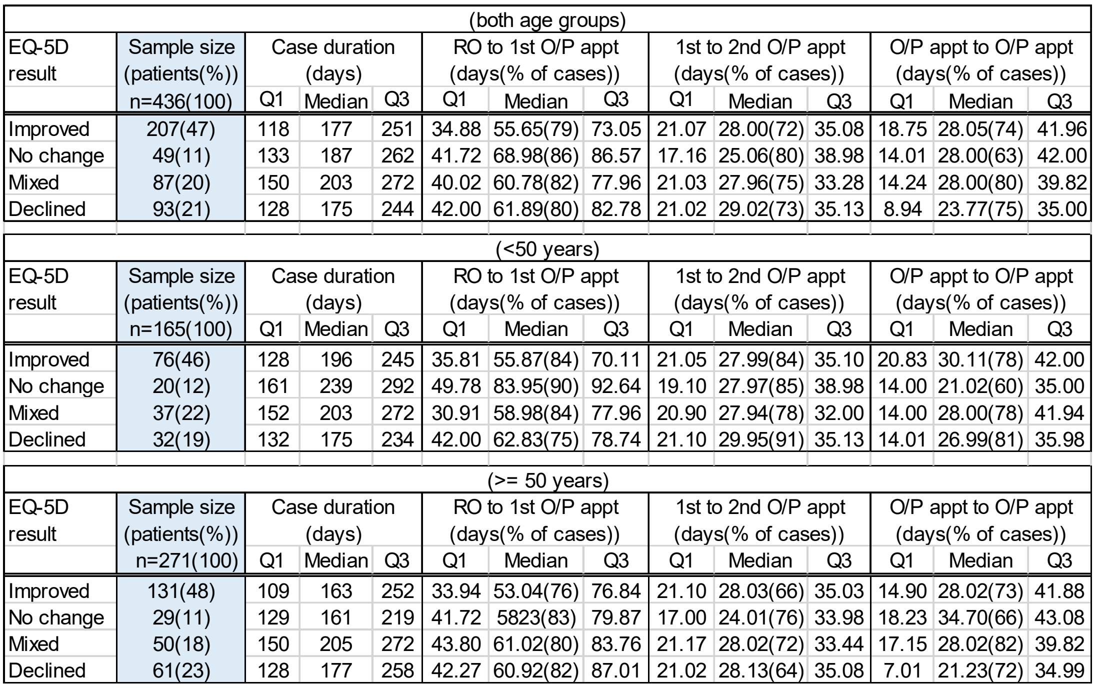
Figure 6.20(a) Proportion of knee pain patients with contact events by sex and health outcome, (b) for those patients, the median number of contact event occurrences



The information presented in Figure 6.20(a)2 shows that almost double the percentage of male patients with declined health outcomes have phone appointments (40%) compared to those with improved health outcomes (24%). However, the information in Figure 6.20(b)2 shows no noticeable difference in the average number of telephone appointments (0.1) for these patients.

Table 6.4 presents information related to the total case duration and time intervals between contact events for knee pain patients grouped by age group and health outcome.

Table 6.4 Knee pain patient time information by age group



(both age groups)

EQ-5D result	Sample size (patients(%))	Case duration (days)			RO to 1st O/P appt (days(% of cases))			1st to 2nd O/P appt (days(% of cases))			O/P appt to O/P appt (days(% of cases))		
	n=436(100)	Q1	Median	Q3	Q1	Median	Q3	Q1	Median	Q3	Q1	Median	Q3
Improved	207(47)	118	177	251	34.88	55.65(79)	73.05	21.07	28.00(72)	35.08	18.75	28.05(74)	41.96
No change	49(11)	133	187	262	41.72	68.98(86)	86.57	17.16	25.06(80)	38.98	14.01	28.00(63)	42.00
Mixed	87(20)	150	203	272	40.02	60.78(82)	77.96	21.03	27.96(75)	33.28	14.24	28.00(80)	39.82
Declined	93(21)	128	175	244	42.00	61.89(80)	82.78	21.02	29.02(73)	35.13	8.94	23.77(75)	35.00

(<50 years)

EQ-5D result	Sample size (patients(%))	Case duration (days)			RO to 1st O/P appt (days(% of cases))			1st to 2nd O/P appt (days(% of cases))			O/P appt to O/P appt (days(% of cases))		
	n=165(100)	Q1	Median	Q3	Q1	Median	Q3	Q1	Median	Q3	Q1	Median	Q3
Improved	76(46)	128	196	245	35.81	55.87(84)	70.11	21.05	27.99(84)	35.10	20.83	30.11(78)	42.00
No change	20(12)	161	239	292	49.78	83.95(90)	92.64	19.10	27.97(85)	38.98	14.00	21.02(60)	35.00
Mixed	37(22)	152	203	272	30.91	58.98(84)	77.96	20.90	27.94(78)	32.00	14.00	28.00(78)	41.94
Declined	32(19)	132	175	234	42.00	62.83(75)	78.74	21.10	29.95(91)	35.13	14.01	26.99(81)	35.98

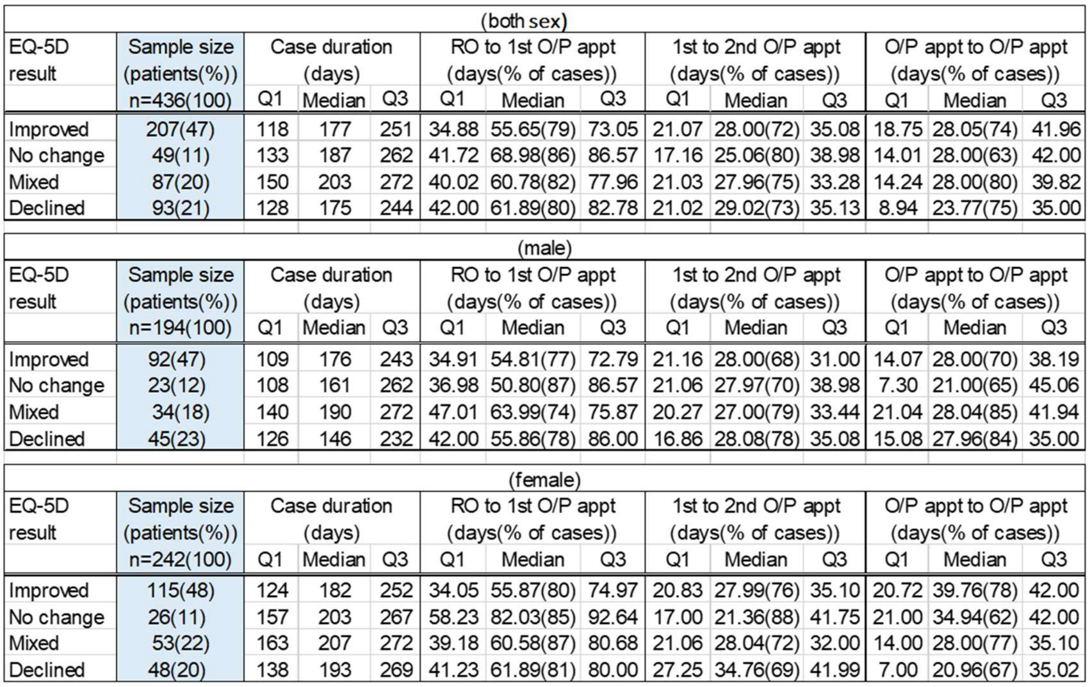
(>= 50 years)

EQ-5D result	Sample size (patients(%))	Case duration (days)			RO to 1st O/P appt (days(% of cases))			1st to 2nd O/P appt (days(% of cases))			O/P appt to O/P appt (days(% of cases))		
	n=271(100)	Q1	Median	Q3	Q1	Median	Q3	Q1	Median	Q3	Q1	Median	Q3
Improved	131(48)	109	163	252	33.94	53.04(76)	76.84	21.10	28.03(66)	35.03	14.90	28.02(73)	41.88
No change	29(11)	129	161	219	41.72	5823(83)	79.87	17.00	24.01(76)	33.98	18.23	34.70(66)	43.08
Mixed	50(18)	150	205	272	43.80	61.02(80)	83.76	21.17	28.02(72)	33.44	17.15	28.02(82)	39.82
Declined	61(23)	128	177	258	42.27	60.92(82)	87.01	21.02	28.13(64)	35.08	7.01	21.23(72)	34.99

On initial examination, there is no obvious difference between the total case durations for patients with improved and declined health outcome (2 days). However, when the event log is divided by age group, younger patients with a declined health outcome have shorter treatment durations (21 days) than patients that improved. For those same patients, the average time intervals between GP referral to first appointment is seven days longer. However, the opposite is seen with the older age group, here patients with a declined health outcome, on average, have longer treatment durations (14 days) than those who improved. Similar to the younger age group, time between GP referral and first appointment is eight days longer for patients with a declined health outcome compared to those who improved. However, for patients in the older age group with a declined health outcome, the time interval between follow-up appointments is, on average, 7 days less.

When the event log is divided by sex, as seen in Table 6.5, the differences between total case durations become more apparent.

Table 6.5 Knee pain patient time information by sex



(both sex)

EQ-5D result	Sample size (patients(%))	Case duration (days)			RO to 1st O/P appt (days(% of cases))			1st to 2nd O/P appt (days(% of cases))			O/P appt to O/P appt (days(% of cases))		
	n=436(100)	Q1	Median	Q3	Q1	Median	Q3	Q1	Median	Q3	Q1	Median	Q3
Improved	207(47)	118	177	251	34.88	55.65(79)	73.05	21.07	28.00(72)	35.08	18.75	28.05(74)	41.96
No change	49(11)	133	187	262	41.72	68.98(86)	86.57	17.16	25.06(80)	38.98	14.01	28.00(63)	42.00
Mixed	87(20)	150	203	272	40.02	60.78(82)	77.96	21.03	27.96(75)	33.28	14.24	28.00(80)	39.82
Declined	93(21)	128	175	244	42.00	61.89(80)	82.78	21.02	29.02(73)	35.13	8.94	23.77(75)	35.00

(male)

EQ-5D result	Sample size (patients(%))	Case duration (days)			RO to 1st O/P appt (days(% of cases))			1st to 2nd O/P appt (days(% of cases))			O/P appt to O/P appt (days(% of cases))		
	n=194(100)	Q1	Median	Q3	Q1	Median	Q3	Q1	Median	Q3	Q1	Median	Q3
Improved	92(47)	109	176	243	34.91	54.81(77)	72.79	21.16	28.00(68)	31.00	14.07	28.00(70)	38.19
No change	23(12)	108	161	262	36.98	50.80(87)	86.57	21.06	27.97(70)	38.98	7.30	21.00(65)	45.06
Mixed	34(18)	140	190	272	47.01	63.99(74)	75.87	20.27	27.00(79)	33.44	21.04	28.04(85)	41.94
Declined	45(23)	126	146	232	42.00	55.86(78)	86.00	16.86	28.08(78)	35.08	15.08	27.96(84)	35.00

(female)

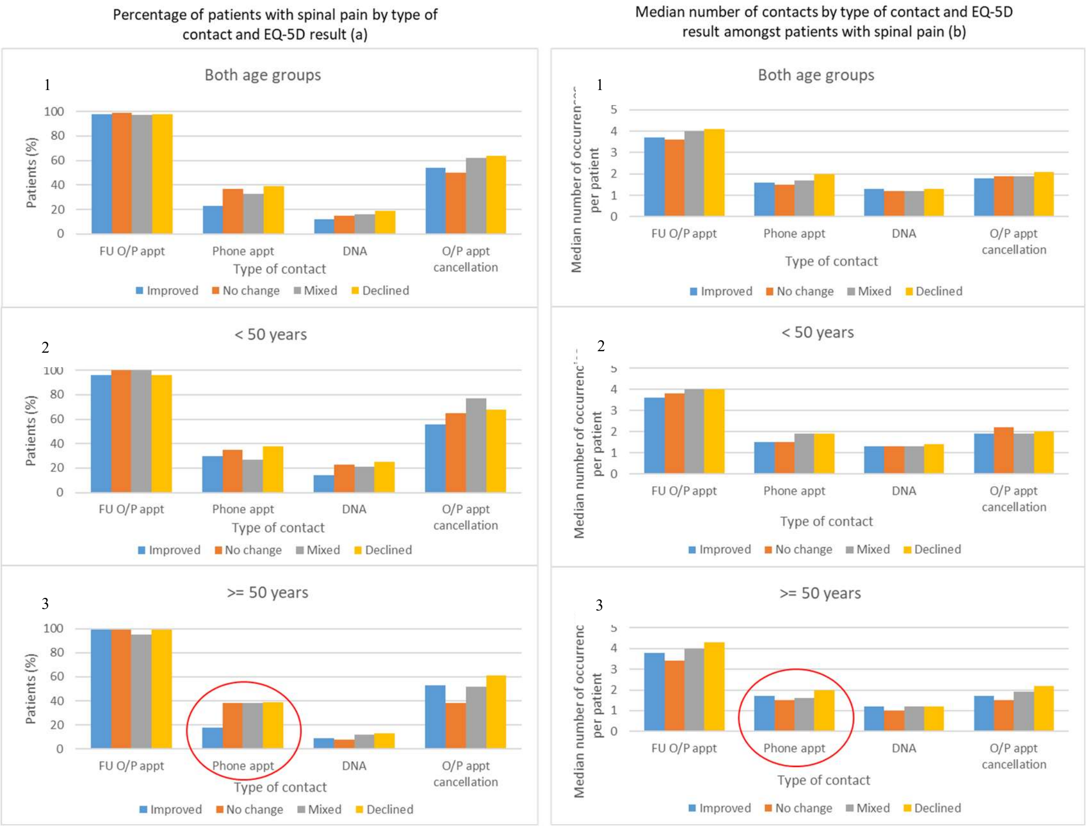
EQ-5D result	Sample size (patients(%))	Case duration (days)			RO to 1st O/P appt (days(% of cases))			1st to 2nd O/P appt (days(% of cases))			O/P appt to O/P appt (days(% of cases))		
	n=242(100)	Q1	Median	Q3	Q1	Median	Q3	Q1	Median	Q3	Q1	Median	Q3
Improved	115(48)	124	182	252	34.05	55.87(80)	74.97	20.83	27.99(76)	35.10	20.72	39.76(78)	42.00
No change	26(11)	157	203	267	58.23	82.03(85)	92.64	17.00	21.36(88)	41.75	21.00	34.94(62)	42.00
Mixed	53(22)	163	207	272	39.18	60.58(87)	80.68	21.06	28.04(72)	32.00	14.00	28.00(77)	35.10
Declined	48(20)	138	193	269	41.23	61.89(81)	80.00	27.25	34.76(69)	41.99	7.00	20.96(67)	35.02

This information shows that on average, male patients are treated for 18 days less than female patients. When considering only male patients, those with a declined health outcome are treated for, on average, 30 days less than those who improved. When the time intervals between appointments are examined, there are no noticeable differences between those with an improved and those with a declined health outcome.

6.4.2.2 Results for spinal pain patients

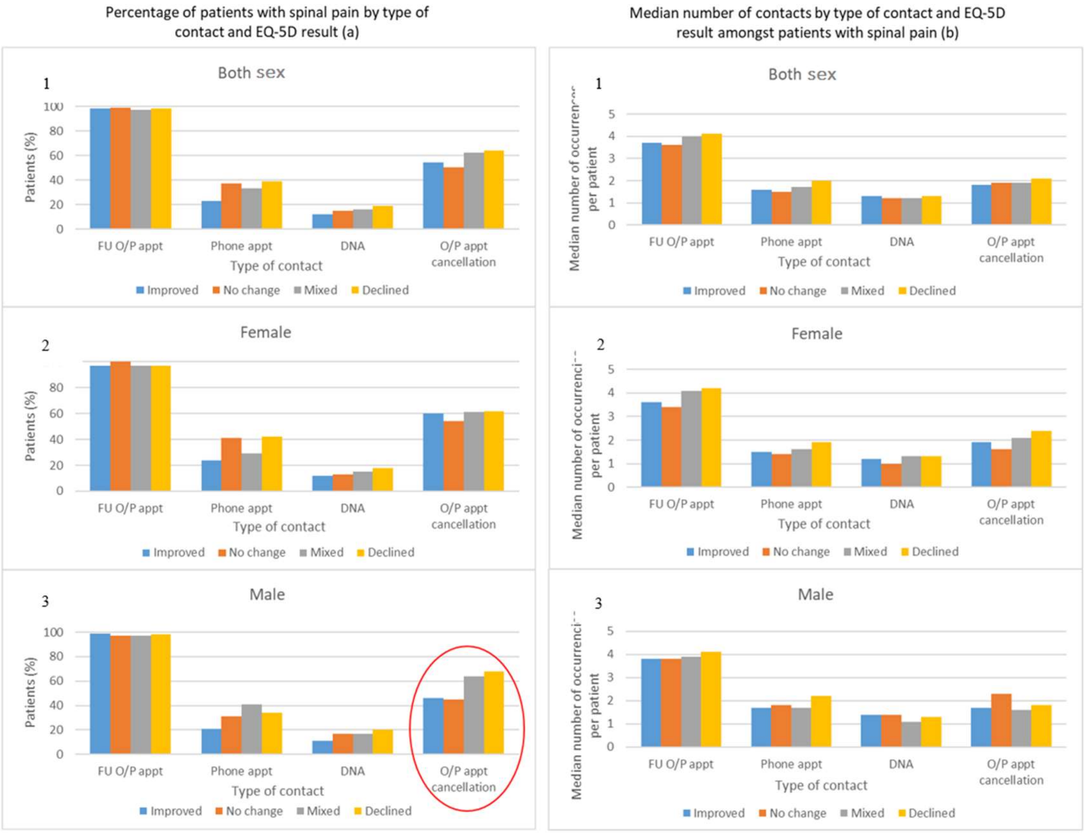
The information presented in Figures 6.21 and 6.22 relates to Figure 6.17. Figures 6.21/22(a) show the percentage of spinal pain patients with each type of contact event, for each of the four health outcomes. Charts 6.21/22(a)1 display this information for all patients and charts 6.21/22(a)2 display the information for patients in the younger age group and males, respectively. Charts 6.21/22(a)3 display the information for patients in the older age group and females, respectively. The charts in Figures 6.21/22(b) correspond to the information shown in (a) by presenting the median number of times that each patient performs a specific content event.

Figure 6.21(a) Proportion of spinal pain patients with contact events by age group and health outcome, (b) for those patients, the median number of event occurrences



Similar to the knee pain patients, it is clear from Figure 6.21(b) that patients with a declined health outcome have, on average, more follow-up appointments than those with an improved health outcome. However, with the spinal pain patients, the most noticeable observation is the difference between the percentage of patients in the older age group having phone appointments (Figure 6.21(a)3). Patients with a declined health outcome have over double the percentage of phone appointments (39%) compared to those with an improved health outcome (18%). In addition, the information in Figure 6.21(b)3 shows that the number of occurrences per patient is also slightly higher in those with a declined health outcome (2.0) compared to those with an improved health outcome (1.7). Figure 6.22 presents the same information for when the dataset is split by sex.

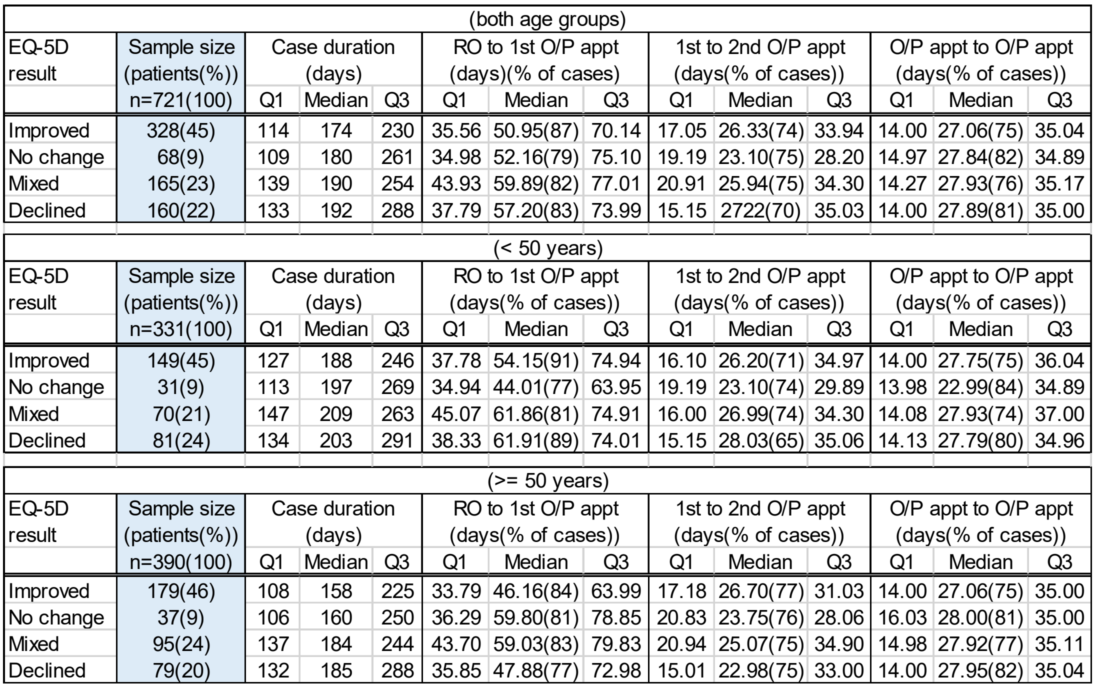
Figure 6.22(a) Proportion of spinal pain patients with contact events by sex and health outcome, (b) for those patients, the median number of contact event occurrences



The information presented in Figure 6.22(a)3 shows that almost a third more male patients with declined health outcomes (68%) cancel their appointments compared to those with improved health outcomes (46%). However, the information in Figure 6.22(b)3 shows no noticeable difference in the average number of times (0.1) that those patients cancelled their outpatient appointments.

The information in Table 6.6 presents the total case durations and time intervals between contact events for spinal pain patients by age group and health outcome.

Table 6.6 Spinal pain patient time information by age group



(both age groups)													
EQ-5D result	Sample size (patients(%))	Case duration (days)			RO to 1st O/P appt (days)(% of cases)			1st to 2nd O/P appt (days(% of cases))			O/P appt to O/P appt (days(% of cases))		
	n=721(100)	Q1	Median	Q3	Q1	Median	Q3	Q1	Median	Q3	Q1	Median	Q3
Improved	328(45)	114	174	230	35.56	50.95(87)	70.14	17.05	26.33(74)	33.94	14.00	27.06(75)	35.04
No change	68(9)	109	180	261	34.98	52.16(79)	75.10	19.19	23.10(75)	28.20	14.97	27.84(82)	34.89
Mixed	165(23)	139	190	254	43.93	59.89(82)	77.01	20.91	25.94(75)	34.30	14.27	27.93(76)	35.17
Declined	160(22)	133	192	288	37.79	57.20(83)	73.99	15.15	2722(70)	35.03	14.00	27.89(81)	35.00

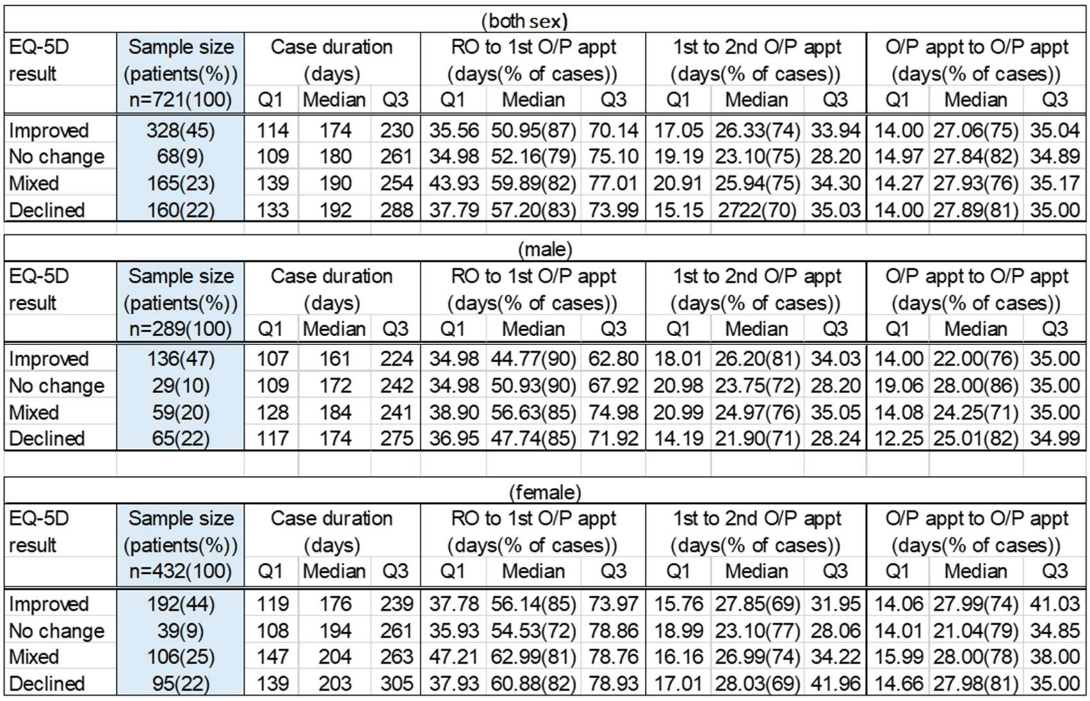
(< 50 years)													
EQ-5D result	Sample size (patients(%))	Case duration (days)			RO to 1st O/P appt (days(% of cases))			1st to 2nd O/P appt (days(% of cases))			O/P appt to O/P appt (days(% of cases))		
	n=331(100)	Q1	Median	Q3	Q1	Median	Q3	Q1	Median	Q3	Q1	Median	Q3
Improved	149(45)	127	188	246	37.78	54.15(91)	74.94	16.10	26.20(71)	34.97	14.00	27.75(75)	36.04
No change	31(9)	113	197	269	34.94	44.01(77)	63.95	19.19	23.10(74)	29.89	13.98	22.99(84)	34.89
Mixed	70(21)	147	209	263	45.07	61.86(81)	74.91	16.00	26.99(74)	34.30	14.08	27.93(74)	37.00
Declined	81(24)	134	203	291	38.33	61.91(89)	74.01	15.15	28.03(65)	35.06	14.13	27.79(80)	34.96

(>= 50 years)													
EQ-5D result	Sample size (patients(%))	Case duration (days)			RO to 1st O/P appt (days(% of cases))			1st to 2nd O/P appt (days(% of cases))			O/P appt to O/P appt (days(% of cases))		
	n=390(100)	Q1	Median	Q3	Q1	Median	Q3	Q1	Median	Q3	Q1	Median	Q3
Improved	179(46)	108	158	225	33.79	46.16(84)	63.99	17.18	26.70(77)	31.03	14.00	27.06(75)	35.00
No change	37(9)	106	160	250	36.29	59.80(81)	78.85	20.83	23.75(76)	28.06	16.03	28.00(81)	35.00
Mixed	95(24)	137	184	244	43.70	59.03(83)	79.83	20.94	25.07(75)	34.90	14.98	27.92(77)	35.11
Declined	79(20)	132	185	288	35.85	47.88(77)	72.98	15.01	22.98(75)	33.00	14.00	27.95(82)	35.04

It is evident from looking at the data for both age groups that patients with declined health outcomes have, on average, longer case durations (18 days), compared to those with improved outcomes. Six out of the 18 days relates to the time between GP referral and first appointment. When the event log is divided by age group, for the older age group, the difference in the average case duration between improved and declined patients becomes more apparent (27 days).

When the event log is divided by sex, as seen in Table 6.7, the difference in the time intervals between contact events for patients with improved and declined health outcomes become more apparent.

Table 6.7 Spinal pain patient delays by sex



(both sex)

EQ-5D result	Sample size (patients(%))	Case duration (days)			RO to 1st O/P appt (days(% of cases))			1st to 2nd O/P appt (days(% of cases))			O/P appt to O/P appt (days(% of cases))		
	n=721(100)	Q1	Median	Q3	Q1	Median	Q3	Q1	Median	Q3	Q1	Median	Q3
Improved	328(45)	114	174	230	35.56	50.95(87)	70.14	17.05	26.33(74)	33.94	14.00	27.06(75)	35.04
No change	68(9)	109	180	261	34.98	52.16(79)	75.10	19.19	23.10(75)	28.20	14.97	27.84(82)	34.89
Mixed	165(23)	139	190	254	43.93	59.89(82)	77.01	20.91	25.94(75)	34.30	14.27	27.93(76)	35.17
Declined	160(22)	133	192	288	37.79	57.20(83)	73.99	15.15	2722(70)	35.03	14.00	27.89(81)	35.00

(male)

EQ-5D result	Sample size (patients(%))	Case duration (days)			RO to 1st O/P appt (days(% of cases))			1st to 2nd O/P appt (days(% of cases))			O/P appt to O/P appt (days(% of cases))		
	n=289(100)	Q1	Median	Q3	Q1	Median	Q3	Q1	Median	Q3	Q1	Median	Q3
Improved	136(47)	107	161	224	34.98	44.77(90)	62.80	18.01	26.20(81)	34.03	14.00	22.00(76)	35.00
No change	29(10)	109	172	242	34.98	50.93(90)	67.92	20.98	23.75(72)	28.20	19.06	28.00(86)	35.00
Mixed	59(20)	128	184	241	38.90	56.63(85)	74.98	20.99	24.97(76)	35.05	14.08	24.25(71)	35.00
Declined	65(22)	117	174	275	36.95	47.74(85)	71.92	14.19	21.90(71)	28.24	12.25	25.01(82)	34.99

(female)

EQ-5D result	Sample size (patients(%))	Case duration (days)			RO to 1st O/P appt (days(% of cases))			1st to 2nd O/P appt (days(% of cases))			O/P appt to O/P appt (days(% of cases))		
	n=432(100)	Q1	Median	Q3	Q1	Median	Q3	Q1	Median	Q3	Q1	Median	Q3
Improved	192(44)	119	176	239	37.78	56.14(85)	73.97	15.76	27.85(69)	31.95	14.06	27.99(74)	41.03
No change	39(9)	108	194	261	35.93	54.53(72)	78.86	18.99	23.10(77)	28.06	14.01	21.04(79)	34.85
Mixed	106(25)	147	204	263	47.21	62.99(81)	78.76	16.16	26.99(74)	34.22	15.99	28.00(78)	38.00
Declined	95(22)	139	203	305	37.93	60.88(82)	78.93	17.01	28.03(69)	41.96	14.66	27.98(81)	35.00

Similar to the results of the knee pain patients, on average male patients are treated for less time than female patients (21.5 days). However, again this does not appear to correlate to worse health outcomes, as 22 percent of male and female patients result in a declined health outcome. The average case duration for female patients with a declined health outcome is 27 days longer than for those who have improved.

6.5 Stage 4: Evaluation

There were two aims to this study. The first was to determine whether process mining techniques could be applied to the MyPathway data, in order to identify possible indicators of a patient health outcome, using PROM data. After a considerable amount of data cleansing and processing, data resulting from matched pairs of EQ-5D questionnaires were used to categorise each patient into one of four health outcome groups. After the generation of 40 event logs (see Section 6.3.5.2 for breakdown), process mining techniques were applied to the MyPathway data and possible indicators of health outcomes were identified. The second aim was to identify where changes could be made to the MyPathway system in order to improve data collection for future process mining studies. Many issues were identified during the data transformation stage and these were discussed with the MyPathway developers and clinical domain experts at STHT before solutions were designed. The work in this chapter has been continuously verified through discussions with the clinical and business experts at STHT and the technical and business experts at ADI. Statistical testing of the results was considered in order to know whether the differences between results was reproducible. However, this did not happen for two reasons: 1) there were too many results to test, all with equal importance, therefore testing would risk too many false positives [381]; and 2) as this is the first study using this kind of data, descriptive statistics should first be ascertained to identify what is potentially clinically important. To determine the clinical importance, the results were presented and discussed with the Clinical Service Manager at STHT. The process mining results were found to be interesting, though it was concluded that due to the small margins of difference and relatively small numbers of patients possible indicators of health outcomes may be down to chance.

No studies could be found within the literature where direct comparison of results could be made. However, a study by Bekkering et al. [382] identified possible indicators of a poor health outcome for low back pain patients receiving physiotherapy treatment. Results showed that the most probable predictor was the duration of the current episode, followed by the time to first physiotherapy appointment. This aligns with the results from the MyPathway study, where time between GP referral and first appointment was also found to be a possible predictor for both spine and knee pain patients. Karmarkar et al. [383] used a Markov model to calculate lifetime costs of knee OA treatment before stratifying patients by race, ethnicity, sex, and socioeconomic status. Results from this study, across all groups, found that delaying treatment may result in worse outcomes. Further support is provided by Chester et al. [384] who carried out a systematic review on predicting response to physiotherapy treatment for MSK shoulder pain, where again, the duration of pain was found to be a possible predictor.

The literature suggests a range of methods for the validation of process models, which often includes measuring the model quality in terms of its replay fitness, precision and generalisation [251]. As explained in Section 3.4.1, replay fitness refers to how well the model can reproduce the observed behaviour from the event log used to discover it. Precision is a measure of how much the discovered model over-estimates the traces in the log and finally, the model can generalise if it can replay behaviour it has not previously seen. These three metrics are typically used to measure a model's quality. However, if the sole purpose of the model is to *visualise* the data generated from a particular event log, then the metrics precision and generalisation do not apply as they relate only to when the model is exposed to behaviour it has not previously seen. The process models in this chapter have been generated for the sole purpose of providing an exact visual representation of the data within each event log. The information extracted from these process models was used to compare different cohorts of patients, in order to identify potential indicators of health outcomes. For this reason, the only quality metric of interest is the fitness metric. A fitness measure of 1.0, represents a complete match between the data in the event log and the discovered process model. All models discovered in this chapter have a fitness measure of 1.0.

6.6 Stage 5: Process improvement

Many of the findings from this work have been used by the development team at ADI to improve their processes. During and after this study, the process mapping diagrams were used in presentations to inform clinicians and other stakeholders.

A significant part of this work was taken up by the cleansing and pre-processing of the MyPathway data. This was done to ensure the viability of the data for use during the mining and analysis stage. The ADI team were aware of some of the issues, though others were identified during this study. All issues found by the author were recorded in an issues register (available in Appendix C) and shared with the developers. Solutions were designed and specified using pseudocode (available in Appendix D). This pseudocode, along with the findings from the research, have been incorporated into the tool by the development team at ADI in order to assist with the improvements detailed in Section 6.3.5.

In addition to the data cleansing activities, many of the data transformations have also been incorporated into the software. For example, all data transformations performed to enrich the event data have been implemented. These include: separating the outpatient appointments into ‘new’ and ‘follow-up’; splitting the types for MyPathway invitations and allocated questionnaires into ‘original’ and ‘reminder’; and creating different types of questionnaires, such as ‘baseline’, ‘pre-treatment’ and ‘discharge’. Making these changes has significantly improved the understandability of the data and allowed for a more accurate representation of patient status at any one point in time.

6.7 Future work

At the time of data extraction, the MyPathway application was in its infancy with many features still work-in-progress. Future work may include repeating the experiments with a higher volume of the STHT data, or repeating the experiments using a different MyPathway dataset. Repeating the analysis would allow for inferential statistics to be applied to these results. In both cases, it may add confidence to the results if the analysis was performed using more than one speciality. However, partly due to the work carried out in this study, there has been significant business and system changes to the MyPathway system. Some of these changes include new triggers for the sending of EQ-5D questionnaires, the recording of key data items such as patient step count and patient diagnosis, and changes to ensure data items for events within a patient referral are linked. Due to these reasons, repeating the analysis would involve a considerably large amount of work.

It is possible for mobile applications such as MyPathway to record sensory data from a wide range of mobile devices, such as step count information from fitness trackers. Other researchers could extend the work in this study by repeating the method using their own event log data. The work from this chapter has been made available to ADI to enable future analysis to be carried out using an adaptation of this method on the MyPathway system.

6.8 Summary

To date, this study represents the only published research which has used the MyPathway data for research purposes. No similar studies have been found using process mining techniques. However, Kaur and Mann [385] developed a mobile application using natural language processing and machine learning algorithms to enable predictive healthcare analytics and optimise the clinical process. Patients used the application to answer questions about their symptoms and were provided with personalised information. In addition, the application was enabled with wireless sensors for the transmission of vital signs to alert the healthcare professionals. A data mining approach was taken by Benis et al. [386] in their study on patterns of patients' interactions with healthcare providers. They hypothesised that the patient's follow-up and health outcomes were influenced by the patient's preferred communication channel. In their results paper [387], 13 communication profiles were identified enabling patient communications to be adapted. Overall, a low level of patient engagement was found, which did not substantially improve with the introduction of technological communication channels. Benis concluded that improving patient engagement must be a combination of technological solutions accompanied by complementary means.

This chapter has described the analysis of the MyPathway dataset. The research questions identified in Section 6.2 have been answered. Results from matched pairs of patient EQ-5D questionnaires were used to stratify patients by health outcome using the PCHC method. Discovery process mining techniques were applied to the event logs and features were identified which appeared to correlate with patient health outcomes. These correlations often became more apparent when the event logs were further divided by sex or age group. As the number of events which can be used to identify indicators of a health outcome are limited and almost always happen in the same order regardless of outcome, there is limited scope to identify indicators based on the sequence of activities. However, process mining techniques have proven to be effective for discovering process models from which key metrics can be compared. Such metrics include the total treatment duration, time interval between contact events and the number of event occurrences for patients. The work in this chapter serves as an exemplar. To add confidence to the results, future studies need to be undertaken using the method described in this chapter before comparing the results. Positive results in this area may suggest further work where monitoring software is embedded into similar mobile applications that contain patient generated data. The identification of possible indicators of health outcomes could then be flagged for early intervention, potentially resulting in a change to the patient's treatment course or pathway.

With an increase in the use of healthcare mobile applications, system developers must be aware of the potential benefits that process mining can bring, especially to research and design systems with data extraction in mind. Equally, when process mining data from mobile health applications, analysts must be aware of the potential issues. Although there are many advantages to NoSQL type architectures (Section 3.1.2), they are open to a host of problems when the data is extracted, which may result in a lack of linkage between data items. For this reason, software developers must be mindful of the additional measures which must be put in place to ensure that data consistency is maintained. It is also common with such databases for there to be no predefined data schema, making it difficult for an analyst to understand the data. A major challenge faced during this study was the considerable amount of data cleansing and processing required to fix missing links before the data could be used. Many of the issues highlighted during this study may significantly extend timescales and make the process of data extraction and transformation more challenging.

The specification and testing of the data extract, along with the writing of the data cleansing rules occupied a large proportion of the time spent on this study. There were many times during Stage 2 of the study, when the challenges could have compromised the success of this research programme. The reason to continue working with the MyPathway data was based on the knowledge that the challenges that arise when working with such data, i.e. from mobile applications, PROMs, NoSQL databases and big data, are naturally a by-product of modern system development, especially within the healthcare sector. The purpose of this study was to explore and test the hypothesis avoiding bias. This was done by performing three separate studies, each using a different dataset generated from different processes and using different technologies that would give rise to a wide variety of challenges and insights. Without experiencing, addressing and documenting these challenges an extensive evaluation of the use of process mining techniques used on healthcare data could not have been performed.

The main contribution for this study is in the field of health informatics, as the insights gained and work carried out have helped to make an existing healthcare information system process aware. This contribution is in direct support of [290], where the need to complement healthcare information systems with the process perspective was identified. Further contribution has been made in this chapter to the process and data science community through the methodological insights.

Chapter 7
Process mining SAIL data for knee pain surgery pathways

7.1 Introduction

This chapter presents the third and final study. An expert-defined, interactive reference model for knee pain surgery is presented. The model is a simplification, where process events have been restricted to make it usable in this research. Using process mining techniques, a cohort of knee OA and knee pain patients from the SAIL data was cleansed and validated by replaying it over the reference model to check for conformance. Once validated, the data was used to generate some episode statistics. The methods described in this study can be repeated by other healthcare researchers using different hospital datasets. The results can be used to inform healthcare professionals by providing estimations on the use of knee pain surgery from both a control-flow and temporal perspective. For researchers repeating this study using their own data, these results may provide a baseline for comparison.

7.2 An illustrative example for knee pain

An illustrative example is provided by following the continued journey of Jack, a 60 year old male office manager. Jack had suffered from bilateral knee pain for several months before making an appointment with his GP. He suffered each day from morning pain and stiffness which lasted for around 15 minutes, before it progressively worsened throughout the day. The GP took Jack's history and performed a physical examination before diagnosing osteoarthritis of the knee. Jack was given a leaflet on self-management which included muscle strengthening exercises and the GP suggested he lose two stones in weight by altering his diet. A prescription for paracetamol was given and Jack was told to return if his symptoms worsened. Twelve months later, Jack returned with worsening pain in both knees and was referred to the orthopaedic department within secondary care. Upon examination, the orthopaedic surgeon recommended Jack undergo arthroscopic knee surgery on both knees to relieve the pain. After a pre-operative appointment, Jack received surgery on his left knee, followed by his right knee eight weeks later. The surgery gave good results and after eight weeks Jack was knee pain free. Unfortunately, eighteen months later the pain returned in Jack's right knee. After a GP examination, Jack was referred back to the orthopaedic department for a second arthroscopy on his right knee.

Six months after the second procedure, the pain gradually returned to both knees. The orthopaedic surgeon recommended Jack undergo total knee replacement (TKR) surgery in both knees, nine months apart. Surgery on Jack's left knee was successful. As planned, Jack returned after nine months for TKR surgery on his right knee. Unfortunately, following this surgery Jack was unable to straighten his knee fully and three months later was re-admitted for manipulation of the knee which worked well. Jack continued to be knee pain free for eighteen years before his right prosthetic joint began to give him an increasing amount of pain. After referral back to the orthopaedic department, Jack received a second knee joint replacement (revision TKR) on his right knee.

Jack's case is used as an example to illustrate a pathway for knee pain surgery and will provide a basis for the following sections in this chapter.

7.3 Stage 1: Planning

During the planning stage, information from the official website [20] was accessed in order to understand the type of data available from the SAIL databank. A data request was submitted and ethical approval was granted on the 02/10/2019 (SAIL project number 0814) via the SAIL Collaboration Review System, which consists of the SAIL Management Team and the Information Governance Review Panel. The data request included all patients, 18 years and over at start of study period, with an MSK event during the study period, which was between 01-01-1997 and 30-09-2017. An MSK event was defined by a code list that included codes in the ICD-10 and CTV-2 and coding formats. All ICD-10 codes in Chapter 13 'Diseases of the MSK system and connective tissue' (M00-99) and all Read code versions 2 and 3 codes below and including 'Musculoskeletal and connective tissue diseases' (N….) and (XaDmf) respectively, including gout codes (C34..) and (X40PQ) respectively were included.

All GP, hospital and emergency data related to clinical events including appointments, referrals, diagnostic test results, diagnoses, prescriptions and family history was included along with sociodemographic and ONS death data. After ethical approval was granted, both technical and business resources from the SAIL team were made available via the SAIL helpdesk.

The aim of the study was to determine whether process mining techniques could be applied to the SAIL data in order to examine surgery pathways for patients with a diagnosis of knee pain including knee OA. Three study-specific research questions were composed towards this aim and as an implementation of the primary research question in Chapter 1, these consisted of the following:

1. Can a knee pain surgery reference model be defined for patients diagnosed with knee pain using process mining techniques?

2. Does the behaviour in the real-life SAIL data conform to the knee pain surgery reference model?

3. Can useful healthcare statistics be generated from the SAIL data for patients with knee pain using process mining techniques?

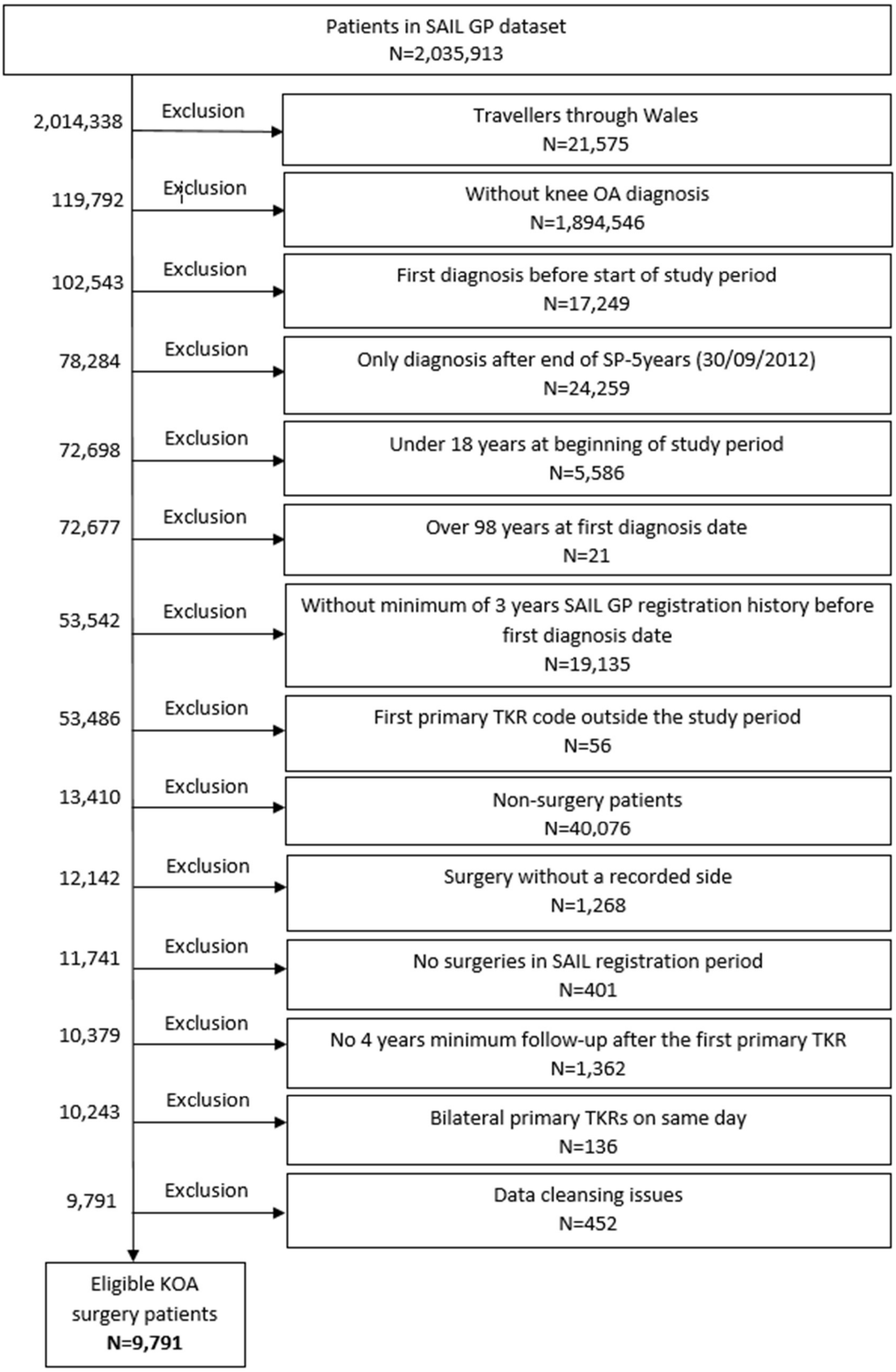
4. What are the strengths and weaknesses of process mining SAIL data?

7.4 Stage 2: Extract, Transform and Load

The SAIL data was accessed as described in Section 4.2.3. To understand the structure and semantics of the data, documentation from the official SAIL website, WIKI and reference library was accessed. The data was also queried and discussions were carried out with the SAIL analysts. Basic validity checks were performed to assess the data quality. For example, ensuring that patients only received treatment whilst they were alive, ensuring ages were within feasible limits and checking that patients were registered at an address in Wales for the duration of their GP SAIL registration.

The exclusion criteria for the study is described using the STROBE diagram presented in Figure 7.1.

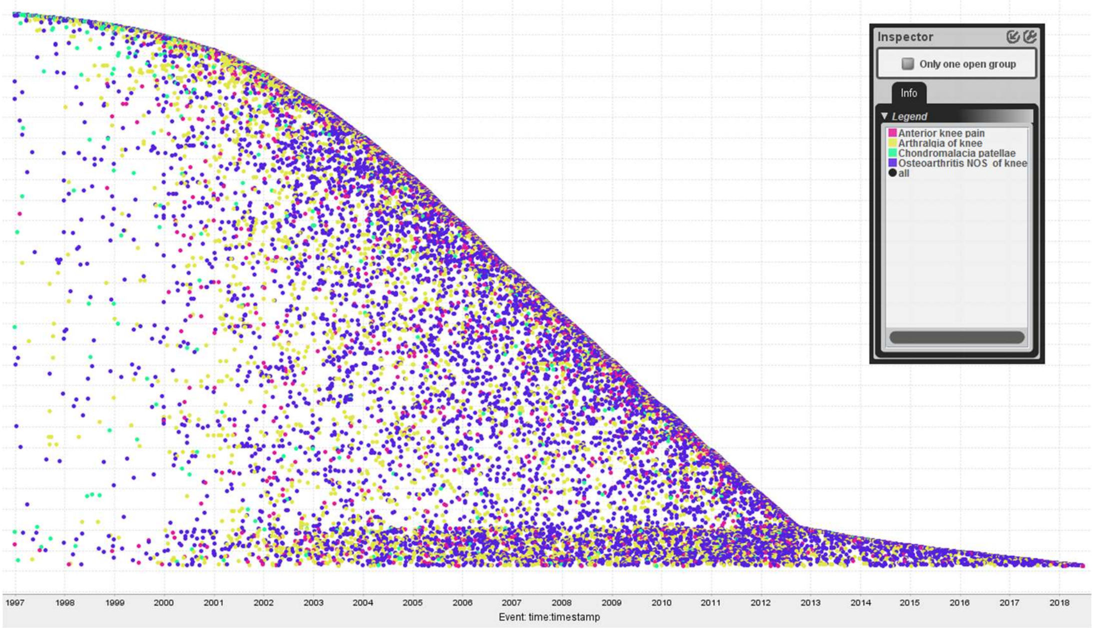
Figure 7.1 STROBE diagram for knee pain surgery cohort


Patients in SAIL GP dataset
N=2,035,913
2,014,338
Exclusion
Travellers through Wales
N=21,575
119,792
Exclusion
Without knee OA diagnosis
N=1,894,546
102,543
Exclusion
First diagnosis before start of study period
N=17,249
78,284
Exclusion
Only diagnosis after end of SP-5years (30/09/2012)
N=24,259
72,698
Exclusion
Under 18 years at beginning of study period
N=5,586
72,677
Exclusion
Over 98 years at first diagnosis date
N=21
53,542
Exclusion
Without minimum of 3 years SAIL GP registration history before first diagnosis date
N=19,135
53,486
Exclusion
First primary TKR code outside the study period
N=56
13,410
Exclusion
Non-surgery patients
N=40,076
12,142
Exclusion
Surgery without a recorded side
N=1,268
11,741
Exclusion
No surgeries in SAIL registration period
N=401
10,379
Exclusion
No 4 years minimum follow-up after the first primary TKR
N=1,362
10,243
Exclusion
Bilateral primary TKRs on same day
N=136
9,791
Exclusion
Data cleansing issues
N=452
Eligible KOA surgery patients
N=9,791

N = number of patients

The SAIL GP dataset consisted of 2,035,913 patients with MSK conditions. Any patient not permanently resident in Wales was excluded to discount people only travelling through. A cohort of knee OA or knee pain patients was selected in order to help design and test the feasibility of the method and the model. Patients without at least one knee OA or knee pain diagnosis between the beginning (01-01-1997) and the end of the study period (30-09-2017) were excluded. Diagnoses were identified by one of the following four CTV-2 codes: 'N05zL' Osteoarthritis NOS of knee; 'N094M' Arthralgia of knee; 'N074.' Chondromalacia patellae; and 'N094W' Anterior knee pain. The choice of codes was guided by the MSK domain experts as they are synonymous with osteoarthritis of the knee. The distribution of these patient diagnoses within the SAIL data is presented in Figure 7.2 using a dotted chart within the ProM framework. The dotted chart is an exploratory tool that helps to identify data quality issues, potential bias or potential patterns of interest. They are a useful tool, especially during the initial stages of exploration where they can be used for the generation of new hypothesis.

Figure 7.2 Visualisation of knee pain diagnoses (n=53,542)



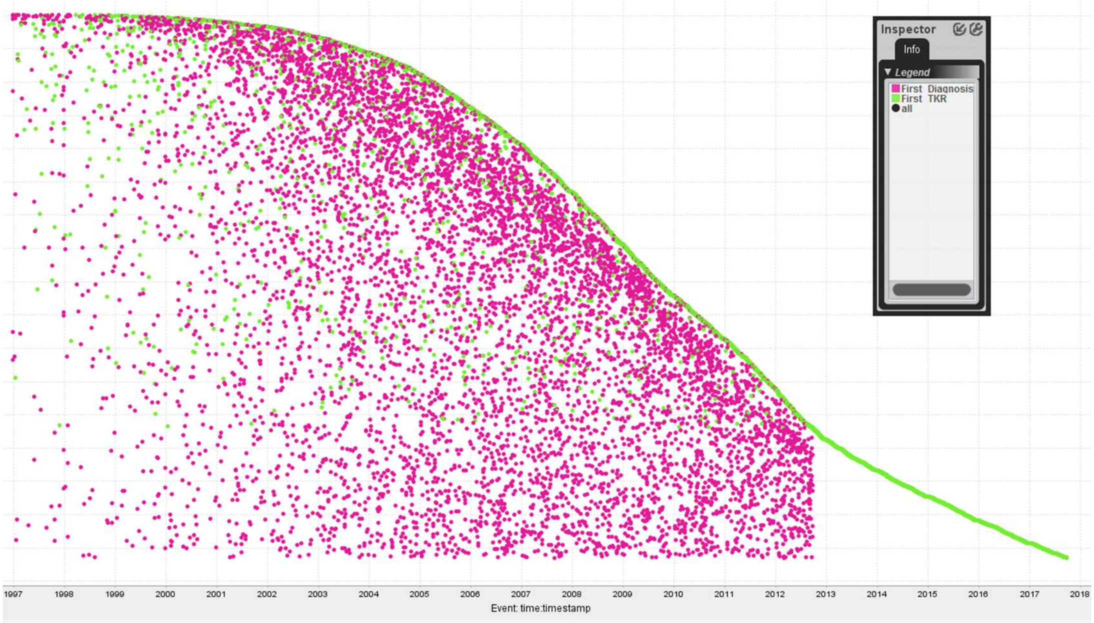
The data is sorted on time of last event and is used to help verify the semantics of the extraction code. To protect patient confidentiality all patient identifiers have been removed from the Y axis. Healthcare data is known to often be of low quality [290].

One of these data quality issues is evident in Figure 7.2, where the chart is less densely populated towards the start of the study period. This is caused by a gradual transition of patient records from paper-based systems to EHRs, resulting in systemic bias within the data. Bias may be introduced into healthcare data by other reasons such as changes to clinical guidelines or the effects on healthcare systems caused by seasonal change.

Figure 7.2 shows a gradual increase, to a linear progression for patients with only one diagnosis totalling approximately 27 percent of the dataset. All patients must have a minimum of five years' worth of follow-up data from the date of their first knee pain diagnosis. When lead and follow-up times before or after a procedure, treatment or diagnosis are included in the selection criteria for a cohort, the number of potential cases is often significantly reduced. This should be taken into account when considering datasets for use in process mining projects. This is evident at a high level by viewing the data in the dotted chart, as the only diagnoses after September 2012 relate to patients with multiple diagnoses with their first before this date.

Within the data, the average, median and upper quartile, for the duration between the patient's first diagnosis and their first primary TKR surgery was 3.3 years, 2.5 years and 5.5 years respectively. This, together with clinical guidance, confirmed that five years' worth of registration data was adequate to allow time after diagnosis for primary TKR surgery. Patients under the age of 18 years at the start of the study period were excluded and to reduce the risk of identification, patients over the age of 98 years at first diagnosis date were also excluded. This was to ensure anonymity for older patients. A visualisation of the durations between a patient's first diagnosis and their first TKR surgery is provided using the dotted chart in Figure 7.3. Again, patient identifiers on the Y axis have been omitted. The event date is displayed on the X axis and all data is sorted by date of last event.

Figure 7.3 Visualisation of durations between first diagnosis and first TKR surgery (n=53,542)



Dotted charts provide an overview of the characteristics of the event log data and help check the correctness of the data extraction code. For example, in Figure 7.3 it is evident that all first diagnoses happen before 30/09/2012, allowing for a minimum of five years' worth of follow-up data before the end of the study period. By visualising the data in this way, it is clear that a large proportion of patients receive their first TKR surgery within approximately two years of their original diagnosis. It is also evident that a small proportion of patients are diagnosed with knee OA after their first TKR, a data quality issue. Possible reasons for this may be the miscoding, retrospective entering or non-recording of original diagnoses. When entering coded information in healthcare information systems administrators and GPs are often presented with a wide choice of codes relating to the same disease or treatment type. Especially when resources are limited and overstretched, it is easy to select an incorrect code. Additionally, information is often retrospectively entered into healthcare information systems, though this is more common practice within secondary, rather than primary care. Finally, there are financial incentives through the QOF for the recording of certain diagnoses such as lung cancer. No QOF codes are associated with OA, which may result in less recordings of the condition.

As seen in the two examples above, the use of dotted charts provides an effective means for visualising the complexities present in healthcare data. In Section 3.5.1, challenge C4 from [290] states that process mining techniques must be able to handle real data generated from healthcare information systems. The two dotted charts above have demonstrated how process mining techniques are useful when working with healthcare data.

All patients must have been registered with a SAIL GP for a minimum of three years prior to their first diagnosis. This rule was to ensure that the patient’s first diagnosis was being used (median time between knee OA diagnoses = 0.99 years (IQR: 0.28 to 2.63)). All first primary TKR surgeries were within the study period. Only patients undergoing knee pain surgery within the study period, with a continuous SAIL GP registration period and where a surgery side was recorded, were included in the analysis. Pathways for the following four types of knee pain surgery were explored, these were: arthroscopy; primary TKR; revision TKR and attention to TKR (AttToTKR). The selection criteria for the OPCS-4 codes was guided by the clinical experts and previously published research which considered different predictors for readmissions following TKR surgery [388].

For the generation of episode statistics, any patient without a minimum of four years’ worth of follow-up data from the date of their first TKR was excluded. This amount of time was chosen to allow for TKR surgery to be performed on the other knee, or for revision surgery to be carried out on the first knee. This number was based on the upper quartile (median duration between first and second TKR surgery = 2.1 years, (IQR: 1.0 to 4.4)), (median duration between TKR surgery and first revision surgery = 2.4 years, (IQR: 1.1 to 4.2)). Surgery events for the 10,379 patients were identified and extracted into an event log using SQL. The exclusion of patients based on the remaining two criteria in Figure 7.1 were carried out during the data transformation stage.

During the data transformation stage, bilateral surgery events were separated into two distinct events, related to right and left side. The meaning of bilateral surgery in this thesis, relates to when a patient has undergone the same surgical procedure on both sides of their body, on the same day. Any patient having bilateral primary TKR surgery was then excluded to be in-line with similar studies [388], [389]. The proportion of bilateral TKR surgery events amounted to 1.1 percent, which is consistent with the number reported in the National Joint Registry [174].

OPCS-4 surgery events were paired and concatenated together with OPCS-4 left or right-sided events. These events were then aggregated to reduce the complexity of the process models. The low-level left and right-sided surgery events were each mapped to one of the four types of high-level surgery events mentioned above, resulting in eight distinct surgery event types. Finally, the event log was loaded into the ProM framework and used with the interactive reference model for knee pain surgery in order to carry out data cleansing activities.

7.5 Stage 3: Mining and analysis

During this study all three types of process mining were used; conformance checking, enhancement and process discovery. Conformance checking was used iteratively to validate and cleanse the dataset. Conformance checking also resulted in the enhancement of the meta-data for the model. Process discovery methods were first used during the initial stage when defining the rules for the reference model. Discovery methods were then used to generate episode statistics using the validated data.

7.5.1 Conformance checking of the SAIL data and enhancement

Conformance checking was performed on the event log data to identify data cleansing issues prior to analysis. As described in Section 3.4.2, conformance checking requires two inputs; an event log and a Petri net model. The event log data, described above, was used in combination with a knee pain surgery reference model. The creation of this model is described in the section below. The MPE plugin within the ProM framework was used to interactively replay the event log data over model. This was an iterative process, with the purpose of identifying any deviations between the data and the model. These deviations were investigated and corrected, in order to arrive at a reusable knee pain surgery reference model and a cleansed and validated dataset ready for analysis. All deviations found during this process were due to one of the following three reasons: 1) a problem with the model; 2) a data cleansing issue; or 3) a permitted rare case (valid exception). The process to create this reference model is described in the following section.

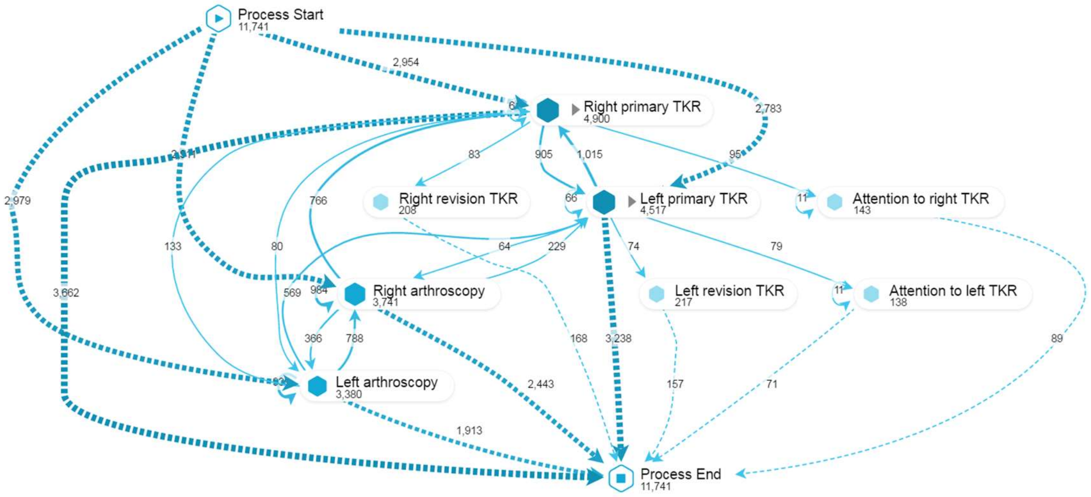
7.5.1.1 Creation of an interactive reference model for knee pain surgery

The process has four steps. The first step was to identify the surgery events.

1) *Identify the sequence and logic for the high-level surgery events*

To help understand the general knee pain surgery process a process model was discovered showing the basic control-flow sequence and frequency of surgery events, this model is presented in Figure 7.4. The Celonis process mining tool was selected to discover the process models. A new project was created and an event log containing 11,741 knee pain patients was imported. Data types were assigned for the case ID, activity and timestamp and a date format was defined. Within the process explorer, the sliders (Figure 4.5) were adjusted to display all eight surgery events types and the top 98 percent of connections, this was to allow for the maximum number of connections that could reasonably be displayed. This feature is particularly useful when wanting to visualise data containing a high level of trace variability, such as healthcare data.

Figure 7.4 Celonis process model for the knee pain surgery pathway



At this point in the process the data had not been cleansed and therefore included errors. However, it provided a good basis for initial discussions with the clinical experts.

2) *Extract a set of rules using the discovered process model and domain expertise*

A meeting was held with a panel of clinical domain experts to establish a set of preliminary rules for the creation of a knee pain surgery reference model. The preliminary rules, extracted from the discovered process model presented in Figure 7.4 were discussed and refined until they were representative of clinical practice. The process model not only acted as a starting point for discussion, it created an additional layer of validation by confirming the thoughts of the experts. The experts thoughts were also evidenced by the behaviour within the data. For example, arthroscopy surgery can be performed multiple times for a patient on the same knee, whereas revision TKR surgery can never be a first surgery for a patient. The following set of rules were defined:

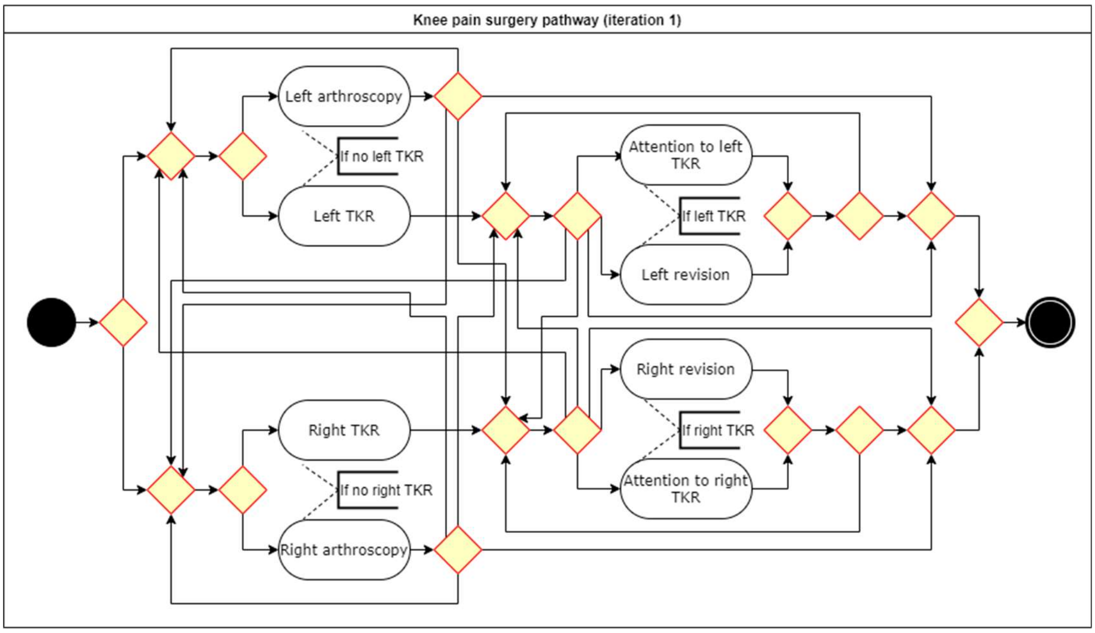
- First event must be an arthroscopy or a primary TKR
- An arthroscopy may lead to a primary TKR
- An arthroscopy cannot occur following a primary TKR in the same knee
- Patients may have multiple arthroscopies

- Patients can have only one primary TKR per knee
- A revision or attention to TKR must only happen following a primary TKR in the same knee
- Patients may have multiple revision or attention to TKRs in the same knee
- Patients may have surgery in one or both knees

3) Create an initial knee pain surgery reference model

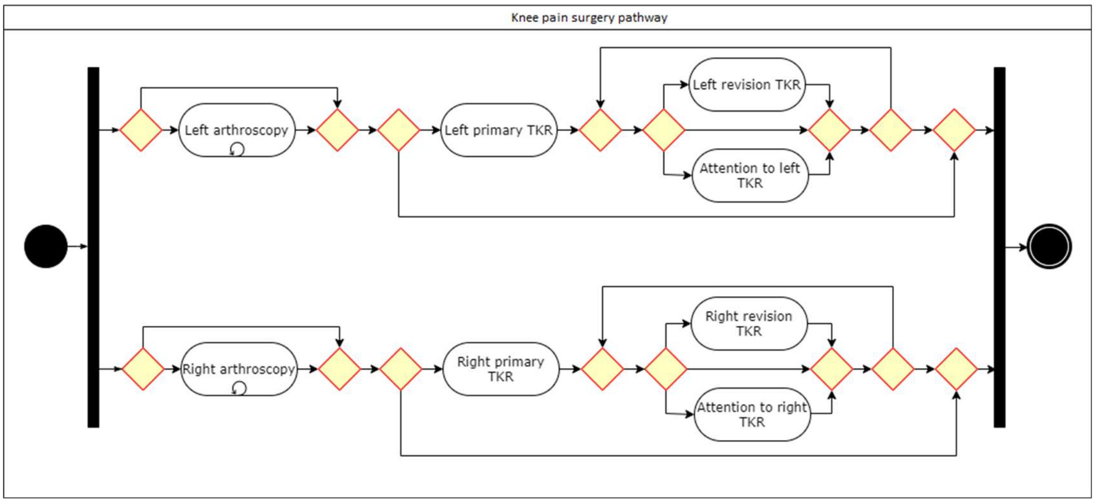
It was important to select a modelling notation that was simple and easy to understand by non-technical people, one that could represent the logic for the knee pain surgery pathway and one where that logic could be directly transferred to the Petri net notation. The main problem was modelling the rules around laterality. UML activity diagrams were chosen for this task. The logic for the knee pain surgery pathway can be modelled in different ways, however, there are advantages and disadvantages to each. The UML activity diagram presented in Figure 7.5 demonstrates one of these ways using exclusive OR gates.

Figure 7.5 UML Activity Diagram for the knee pain surgery pathway using XOR gates



This model is complex and requires the use of guard conditions, which are attached to the eight activities, to accurately model the pathway. However, this results in an extremely precise model, as it does not allow for non-conformant behaviour (any behaviour it has not previously seen). The model presented in Figure 7.6 overcomes this issue by using two logical AND gates.

Figure 7.6 UML Activity Diagram for the knee pain surgery pathway using AND gates



This model is far simpler and does not require the use of guard conditions. However, it does lack in precision, allowing for patients with no behaviour. As event logs generated from EHR data will never contain patients with zero behaviour, it was decided to use this model. It also had the added benefits of being easier to use during discussions with the clinical experts and was simpler to translate into Petri net form. The activity diagram was approved by the clinical experts before it was translated into a Petri net model for use with conformance checking in ProM.

4) Conversion of the activity diagram into a Petri net model

As stated in Section 3.4.5, when Petri net models are used for conformance checking purposes they must be of the type PNML. In this study, to demonstrate different challenges and advantages, two methods for creating a Petri net model were explored. The first method made the use of ProM plugins in order to generate a model using synthetic data and the second method used modelling software.

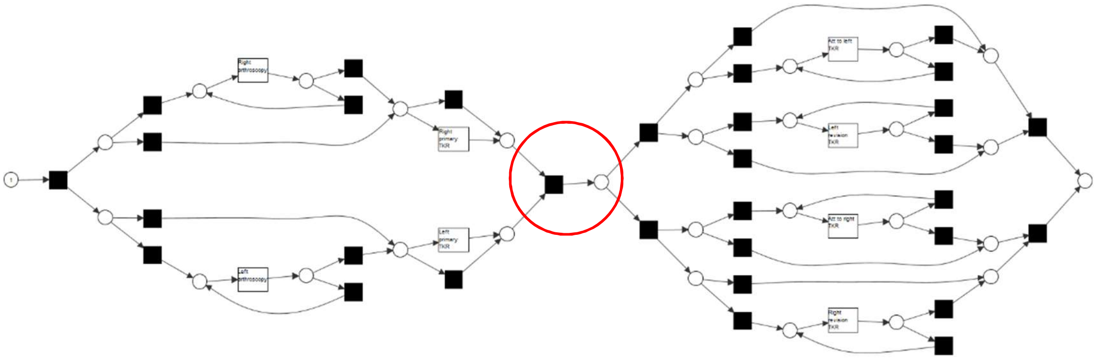
1) Creating a Petri net model in ProM using synthetic data

Microsoft Excel was used to produce an event log using synthetic data. This event log had the standard three columns; case ID, activity and timestamp and consisted of 2,158 events over 388 cases. The following two steps were taken to create the model:

1. Discover Petri net using 'Mine Petri net with Inductive Miner' plugin

The 'Mine Petri net with Inductive Miner' plugin was used to generate the model using the synthetic event log data and the settings described in Section 4.1.3.3. An iterative approach was taken, where a single knee was initially modelled. When adding data for the second knee, left and right pathways incorrectly merged after the two primary TKR events (see Figure 7.7). This resulted in vital information as to which knee, if any, had previously undergone the TKR surgery being lost. Creating a Petri net model that included sidedness using this plugin proved to be extremely difficult. It took many attempts, including conversations with members on the ProM forum [390], until the conclusion was reached that the plugin could not model the data correctly.

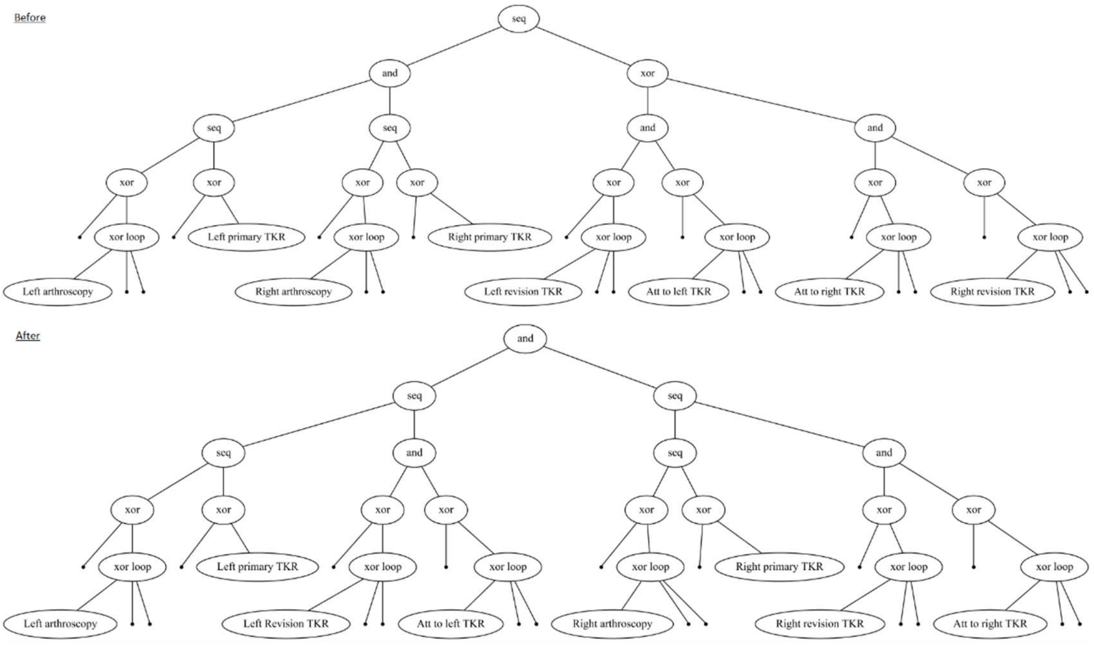
Figure 7.7 Petri net for reference model with incorrect join



2. Discover process tree using 'Mine process tree with Inductive Miner' plugin

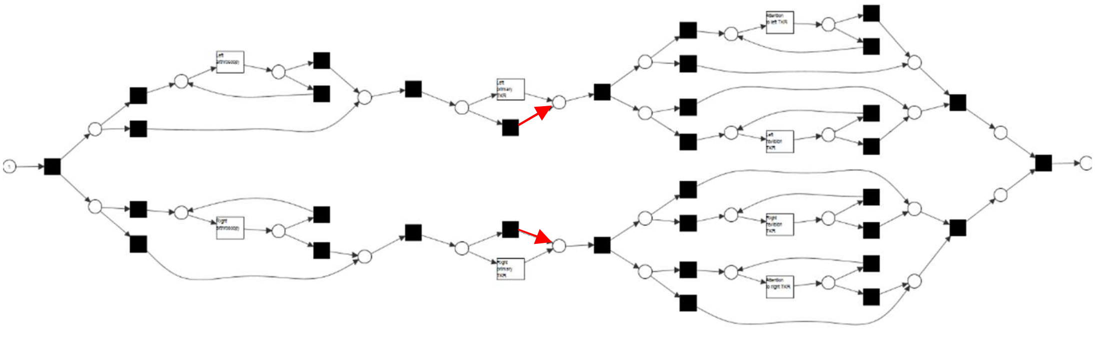
Editing process trees using a graphical editor is far simpler than editing process models. Therefore to overcome this problem, the 'Mine process tree with Inductive Miner' plugin was applied to the same synthetic data, using the settings described in Section 4.1.3.3. The results before and after editing are presented in Figure 7.8.

Figure 7.8 Process tree before and after editing



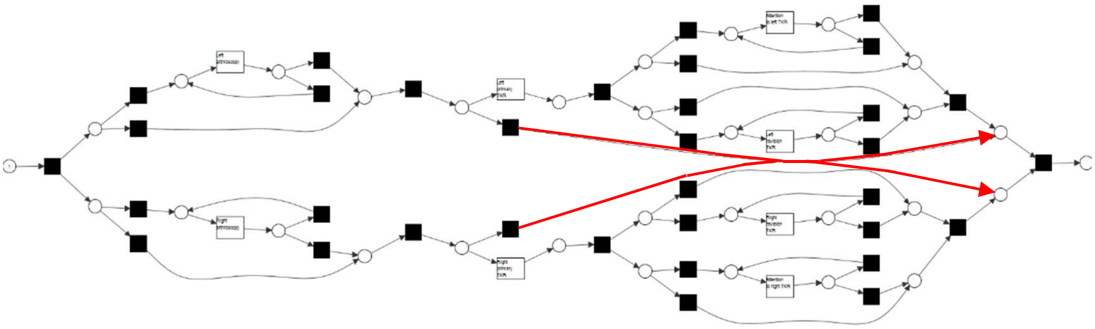
The 'Convert Process Tree to Petri Net' plugin was applied to the edited process tree in order to produce the Petri net model presented in Figure 7.9.

Figure 7.9 Petri net model after editing the process tree



It is clear that the problem presented in Figure 7.7 had been resolved. However, it was not possible to fully correct the model using the 'Mine process tree with Inductive Miner' plugin. Therefore, to allow for only patients who had undergone primary TKR surgery to proceed to revision or attention to TKR surgery (shown in red) it was necessary to further edit the file. Finally, the PNML source code was edited to redirect the two arcs highlighted in red in Figure 7.10.

Figure 7.10 Petri net model after editing

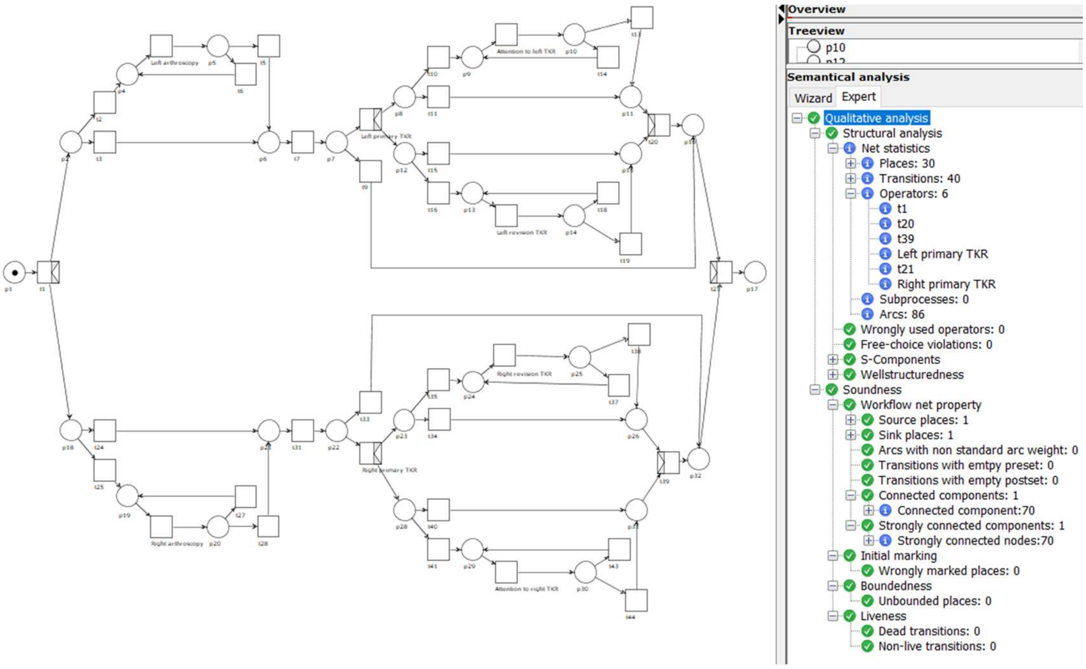


This Petri net model is suitable for use, along with the real patient event log data, in ProM for conformance checking purposes.

2) Creating a Petri net model using modelling software

The Petri net modelling software WoPeD was used to create the model presented in Figure 7.11. This model has been created using the logic specified using the UML activity diagram in Figure 7.6.

Figure 7.11 Petri net model developed using WoPeD



```
Qualitative analysis
  Structural analysis
    Net statistics
      Places: 30
      Transitions: 40
      Operators: 6
        t1
        t20
        t39
        Left primary TKR
        t21
        Right primary TKR
      Subprocesses: 0
      Arcs: 86
    Wrongly used operators: 0
    Free-choice violations: 0
    S-Components
    Wellstructuredness
  Soundness
    Workflow net property
      Source places: 1
      Sink places: 1
      Arcs with non standard arc weight: 0
      Transitions with emtpy preset: 0
      Transitions with empty postset: 0
      Connected components: 1
        Connected component:70
      Strongly connected components: 1
        Strongly connected nodes:70
    Initial marking
      Wrongly marked places: 0
    Boundedness
      Unbounded places: 0
    Liveness
      Dead transitions: 0
      Non-live transitions: 0
```

The window displayed to the right of Figure 7.11 presents a detailed breakdown of the semantical analysis showing the quality of the model. The information shows that the model has achieved 100 percent in both structural analysis and soundness qualities. The structural analysis reports on: the number of and lists the individual model elements; operators that have been incorrectly used; possible logical XOR violations; the number of any sub-processes with their elements; and any place and transition pairs that are not connected by at least two disjointed paths. Soundness reports on: the properties for a workflow net; the presence of a correct initial marking with a token in the sink place; the boundedness, which ensures that places cannot hold multiple tokens at one time; and the liveness, which ensures that all transitions have the ability to fire or be executed.

Unfortunately, the WoPeD software does not support silent transitions (shown in black in Figure 7.10). Therefore, when the model and the data are imported for conformance checking in ProM, the high number of un-named transitions on the model cause an overall fitness score of zero percent. This is easily resolved in ProM by running the 'Configure Visibility of Transitions' plugin [391] against the Petri net model prior to conformance checking.

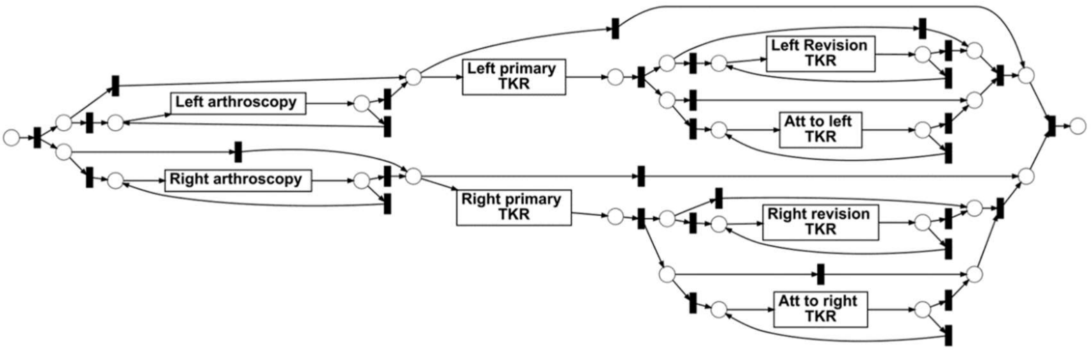
7.5.1.2 Validation using the interactive reference model

Validation of the model and the data was performed using conformance checking techniques. The data from the event log, created in Section 7.4 and the Petri net model created using the WoPeD software were used as input for the 'Multi-perspective Process Explorer' plugin (see Section 4.1.3.3) within the ProM framework. Violations between the two were identified using the Fitness view (Figure 7.13) and investigated using the Trace view (Figure 7.14). A formal meeting was held with the panel of clinical experts. The meeting had two objectives, these were: 1) to present the interactive model, in order to obtain feedback on its use from a clinical perspective; and 2) to address the list of violations, in order to improve the reference model and identify data cleansing issues. The validation process is described using the following six steps.

1) Perform conformance checking using the MPE plugin

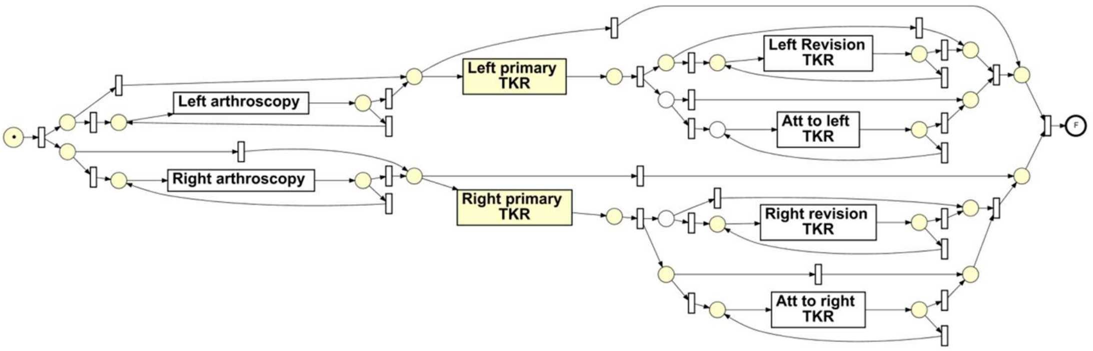
By replaying the event log data over the model within the MPE, it is possible to view the data in five different modes. Figure 7.12 presents the event log data in model mode.

Figure 7.12 Event log data presented in model mode



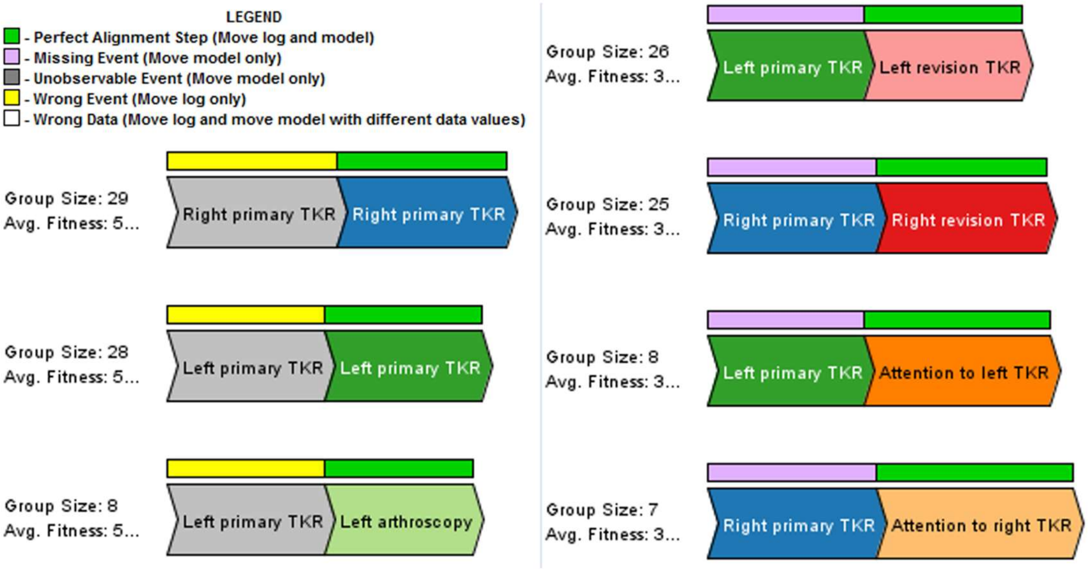
The statistics generated using this mode include the total number of patients in the event log (n=11,741) and the number of events (n=20,326). The data viewed in fitness mode is presented in Figure 7.13.

Figure 7.13 Fitness view, first iteration



The statistics generated in fitness mode produced an average (mean) fitness score of 98.5 percent, with 2.5 percent event violations. These violations comprised of 304 wrong events and 203 missing events (see Section 4.1.3.3). By viewing the data in trace mode, it was possible to perform a detailed analysis of these event violations, an example of this is presented in Figure 7.14.

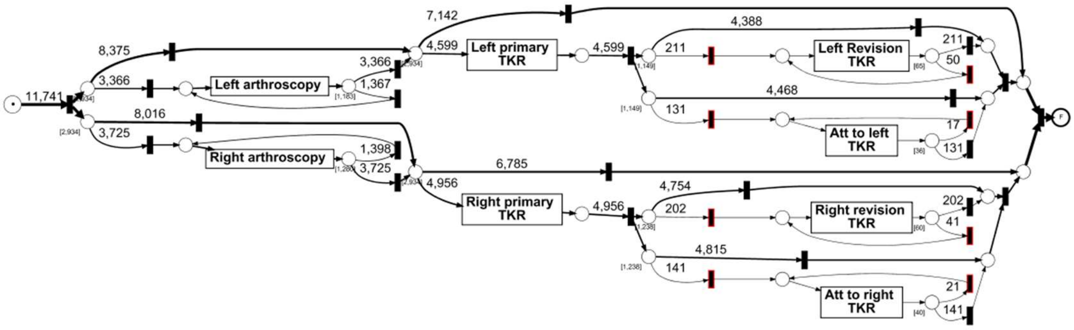
Figure 7.14 MPE Trace view: wrong and missing events



Within the trace view, variants can be sorted by frequency, fitness or length. To the left of each variant is displayed the group size and the average fitness. Above each event is a coloured bar that indicates the status of each event within the variant, the five statuses are displayed in the legend. When working with, often low quality, healthcare data this is a useful feature as it can quickly identify the majority of the problems. For example in the figure above, most of the data quality issues are caused by the incorrect recording of either the laterality or the operation type, as a person cannot have more than one primary knee replacement on the same knee.

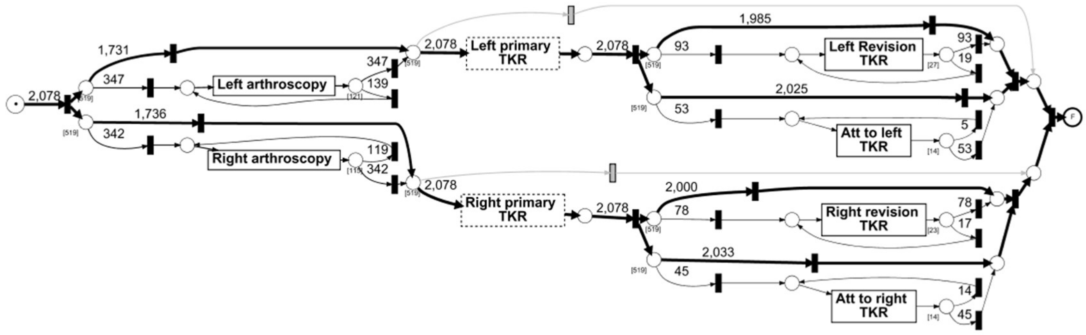
A list of patients is available for each trace variant, allowing for further investigation using the base data, however this list could not be exported using the MPE. Therefore, for this task only, the process mining tool DISCO was used to select and export the list of patients to be used in Step 2. Figure 7.15 presents the event log data in data discovery mode.

Figure 7.15 MPE Data discovery view



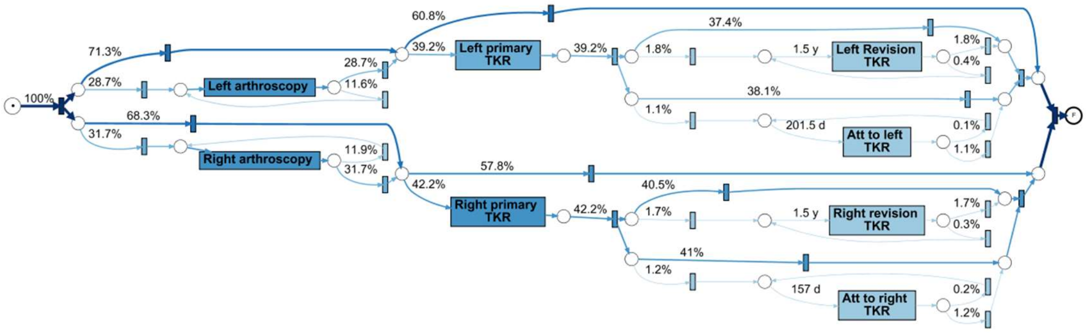
Viewing the model in this mode allows for the number of patients travelling through the different routes to be visualised. The information displayed in Figure 7.15 shows that out of a total of 11,741 patients, 3,366 received a left arthroscopy and 3,725 received a right arthroscopy. 4,599 patients received a left primary TKR and from those, 211 had a left revision TKR and 131 had attention to their left TKR. Violations are highlighted in red against eight silent transitions. Figure 7.16 provides an example of how the event log data can be filtered.

Figure 7.16 Data discovery view filtered by patients receiving TKR surgery on both knees



Here, the two primary TKR event types were selected and a logical AND condition was applied to the data, therefore only data for patients receiving both left and right TKR surgery is displayed. This is a useful feature when wanting to explore differences between patients having single or double sided surgery. Figure 7.17 presents the percentage of patients traversing the model, showing the median time interval between specific points.

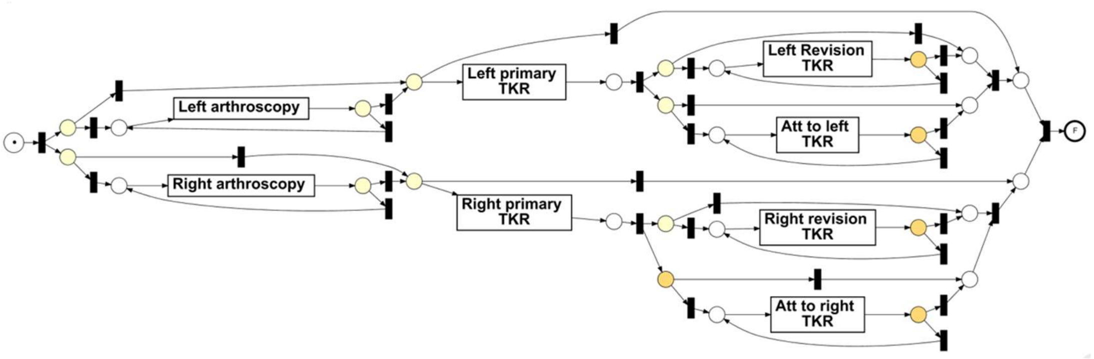
Figure 7.17 Performance view



Unfortunately, in this mode time is not consistently displayed with the same units, making comparison awkward. However, using this view, it is easy to obtain information such as: attention to primary TKR is typically performed much sooner than revision TKR surgery. The MPE software does not provide time intervals between all events. For this reason, the Celonis software was used to discover process models in order to generate some episode statistics.

Figure 7.18 presents the event log data when viewed in precision mode. As described in Section 3.4.1, precision is a process mining metric used to assess the quality of a process model. Precision is high if the model mostly allows for behaviour only seen in the event log.

Figure 7.18 Precision mode



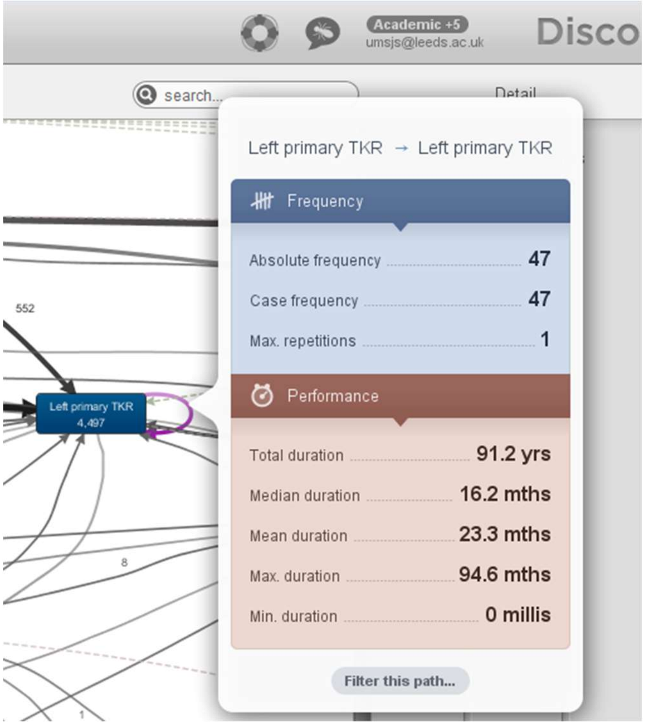
A precision score of 85.2 percent was calculated, which is low, though as previously discussed, precision is not important within this study.

2) Prepare list of violations

An Excel list was compiled containing all patient events from the list of event violations identified using the MPE trace view (see Figure 7.14). Each type of missing or wrong violation (e.g. 'Left primary TKR' → 'Left primary TKR') had an associated list of patients. As stated above, it was not possible to export the list of patients out of the MPE software within ProM, therefore DISCO was used for this purpose. To do this, the following steps were carried out:

1. The same event log was imported into DISCO to create a process model.
2. Within the process model the connection relating to the trace containing the violation was selected and filtered, specifying the path using either the 'directly followed', 'eventually followed', 'never directly followed' or 'never eventually followed' condition, as demonstrated in Figure 7.19.

Figure 7.19 Example of a connection containing violations using Disco



Left primary TKR → Left primary TKR

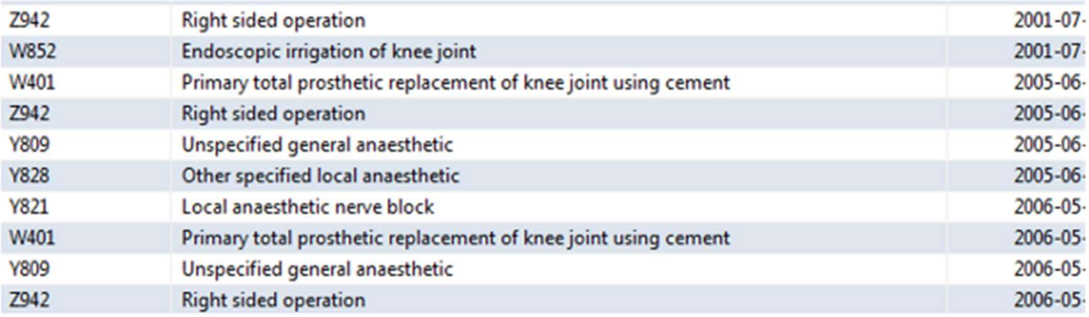
Absolute frequency 47
Case frequency 47
Max. repetitions 1

Total duration 91.2 yrs
Median duration 16.2 mths
Mean duration 23.3 mths
Max. duration 94.6 mths
Min. duration 0 millis

3. Within the filtered process model, a list of patients was exported as an event log in csv format.

4. Patients were added to the master violations spreadsheet, along with contextual examples, using data from the base tables, as seen in Figure 7.20.

Figure 7.20 Query results showing patient surgery data



Z942	Right sided operation	2001-07-
W852	Endoscopic irrigation of knee joint	2001-07-
W401	Primary total prosthetic replacement of knee joint using cement	2005-06-
Z942	Right sided operation	2005-06-
Y809	Unspecified general anaesthetic	2005-06-
Y828	Other specified local anaesthetic	2005-06-
Y821	Local anaesthetic nerve block	2006-05-
W401	Primary total prosthetic replacement of knee joint using cement	2006-05-
Y809	Unspecified general anaesthetic	2006-05-
Z942	Right sided operation	2006-05-

A snapshot of the base table data for a specific patient is presented in Figure 7.20. Information relating to the patient identifier and day have been removed. This data states that the patient has undergone two primary TKR surgeries (W401) on the right knee almost one year apart. Viewing the data at this level of detail highlighted other low-level knee pain surgery codes that may have been overlooked during the original data extract.

5. New surgery types were listed for discussion with the clinical experts.

3) Present interactive reference model to clinical experts

After all information had been collected, a formal walkthrough meeting with the clinical experts took place. The event log data was presented using the five views described in step 1. The experts were impressed with the amount of flexibility, in terms of filtering and the speed at which the event log data could be visualised and interactively explored. The walkthrough provided the clinical experts with confidence in the validity of the data.

4) Investigate event violations with clinical experts

The list of event violations created in step 2 was investigated. Groups of similar cases were discussed to determine whether the violation was due to a problem with the model or a data cleansing issue. A decision was made to change to the meta-data for model. This included the addition of three surgery codes in the original data extract. Two arthroscopy codes (‘W711’, ‘W833’) and one attention to TKR code (‘W913’) were added.

All other changes related to data cleansing issues, often where patients were recorded as having surgeries on the same side in the wrong order, or revision TKR surgery without first having primary TKR surgery. In some cases, it was evident that the wrong side had been coded against the surgery. Two hundred and ninety-three of these violations were wrong events, all of which were illegal according to the rules defined in the previous section. Missing events totalled 217, of which only ten were allowed. These ten events all related to variants where the patient had undergone revision TKR surgery directly after a particular type of primary TKR surgery. The appropriate action was entered against all patient violations, this was either to remove the patient or to amend the event in the event log.

5) Iteration 2, amend the event log and model enhancement

Event data was deleted or amended according to the actions arising from step 2, resulting in the removal of 452 patients from the event log. The meta-data for the model was amended with the addition of the three low-level surgery codes. Events related to these three surgery codes were extracted from the data and added to the event log.

6) Iteration 2, conformance checking

The new event log was replayed over the interactive reference model using the MPE and an average fitness score of 100 percent achieved. The event log data could then be used for analysis, in the knowledge that it had passed the highest rigorous validation process.

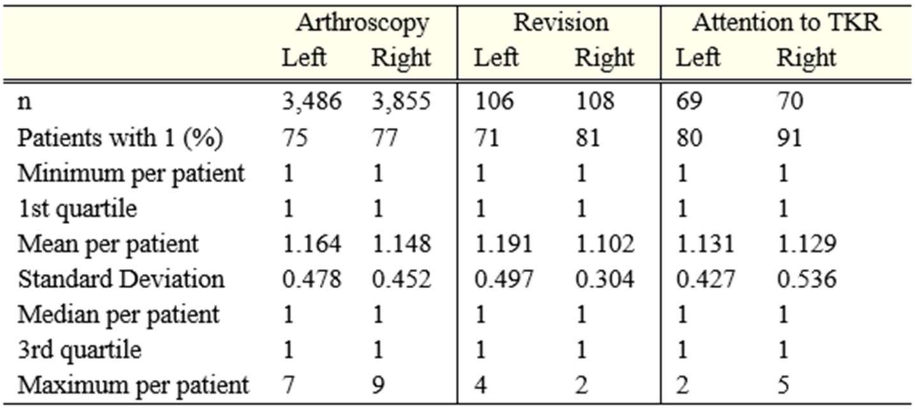
The interactive reference model was primarily created for use in this study, though by following this method, it may be used by other researchers to validate their own knee pain surgery data.

7.5.2 Process discovery and analytics

The process mining tool Celonis was used to perform process discovery using the validated event log data. At this stage, the purpose of process discovery was to generate knee pain surgery pathways containing patient surgery frequency and temporal information. Useful analytics, in the form of episode statistics were extracted from the process models. These statistics and measures may be used by other researchers and medical professionals when considering knee pain surgery.

The clinical experts were interested in the statistics for all repeating surgery types, this information is presented in Table 7.1.

Table 7.1 Repeating surgery type statistics



	Arthroscopy		Revision		Attention to TKR	
	Left	Right	Left	Right	Left	Right
n	3,486	3,855	106	108	69	70
Patients with 1 (%)	75	77	71	81	80	91
Minimum per patient	1	1	1	1	1	1
1st quartile	1	1	1	1	1	1
Mean per patient	1.164	1.148	1.191	1.102	1.131	1.129
Standard Deviation	0.478	0.452	0.497	0.304	0.427	0.536
Median per patient	1	1	1	1	1	1
3rd quartile	1	1	1	1	1	1
Maximum per patient	7	9	4	2	2	5

Results from the SAIL data were compared against results from a study by Espinosa et al. [389] who used Swedish registry data from patients undergoing total hip or knee replacement surgery due to OA. Figures from this study for TKR surgery are presented in Table 7.2 [389].

Table 7.2 Frequencies from the Swedish Knee Arthroplasty Register data for patients undergoing single and subsequent knee replacement surgery

First TKR		No further	Second primary TKR	
	n	Primary TKR	Right	Left
Right	65,895	48,248 (73)	-	17,647 (27)
Left	56,744	40,069 (71)	16,675 (29)	-

n = patients, (%)

In this table the information for patients undergoing hip replacement surgery have been removed, leaving a total of 122,639 patients for comparison purposes. The Celonis process mining software was used to discover process models from the SAIL event log data. Information from these models was collected and presented in tables in order to analyse the frequency of surgeries. Figure 7.21 demonstrates how easily the event log can be filtered.

Figure 7.21 Selection of patient cohort using Celonis

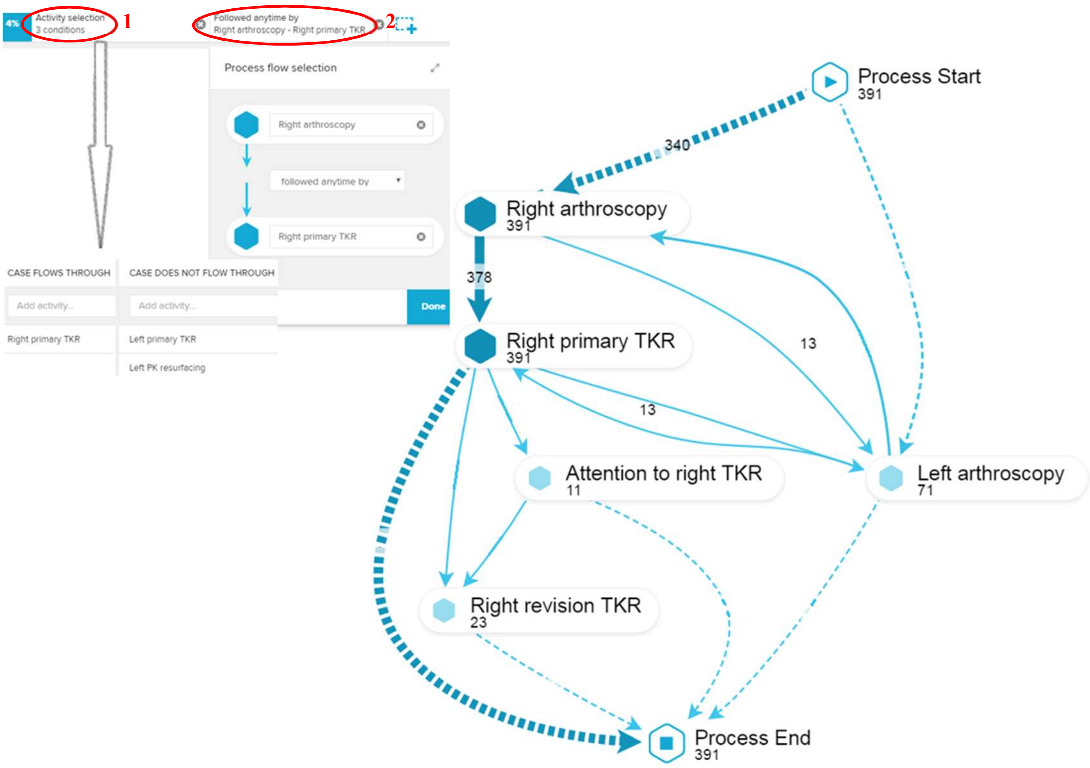
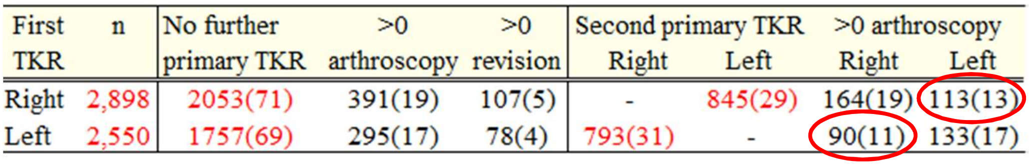


Figure 7.21 presents a process model for patients having arthroscopy surgery on their right knee, followed by *single* knee replacement surgery on that same knee. To select these patients, the event log data was: 1) filtered by activities that all cases either must or must not flow through; 2) filtered by the sequence of those activities, where the nature of the relationship was: ‘directly follows’, ‘follows’, ‘not directly follows’ or ‘not follows’. This method was used to collect the data in Table 7.3 where the number of patients with single and subsequent knee replacement surgery is presented. Values in red can be directly compared with those from Table 7.2.

Table 7.3 Number of SAIL patients who underwent single and subsequent knee replacement surgery



First	n	No further	>0	>0	Second primary TKR		>0 arthroscopy	
TKR		primary TKR	arthroscopy	revision	Right	Left	Right	Left
Right	2,898	2053(71)	391(19)	107(5)	-	845(29)	164(19)	113(13)
Left	2,550	1757(69)	295(17)	78(4)	793(31)	-	90(11)	133(17)

n = patients, (%)

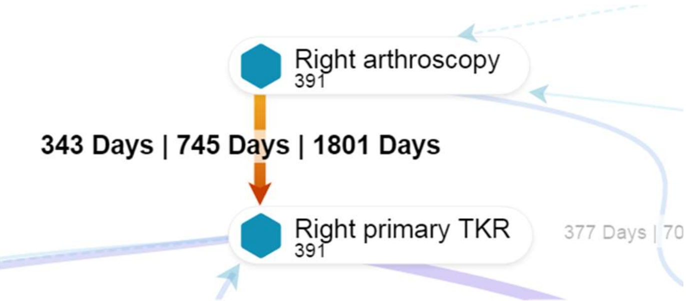
The total number of patients receiving primary TKR surgery is 5,448. In addition to the minimum requirement of three years continuous SAIL registration before first diagnosis and five years follow-up after, patients had a minimum of four years follow-up after their first primary TKR surgery. This number was based on the results of the following calculations: *days between primary TKR surgeries (Q3=1,533) and days between first primary TKR and first TKR revision surgeries (Q3=1,539)*. The results in tables 7.2 and 7.3 both show that patients are at a slightly increased risk of second TKR surgery when the first surgery is on the left side.

In Table 7.3, the data circled in red shows a lower percentage of patients (6%) receive arthroscopy surgery on their second knee. This may be because arthroscopies are generally performed on younger patients to delay the date of the joint replacement. If previous TKR surgery has taken place, the patient is more likely to be suffering from severe OA which will often require a joint replacement. The following caveats were added by an orthopaedic surgeon: 1) As patient and surgeon may be aware that previous arthroscopy surgery was unsuccessful, they may decide to immediately proceed with TKR surgery on the second knee.

2) Regulations and guidance from NICE relating arthroscopic knee washout, with or without debridement, for the treatment of OA changed in 2008 [392]. Previously, arthroscopies were practised more empirically and many patients received unnecessary procedures. The rules surrounding arthroscopies are now more stringent, stating that surgery should not offered unless there is clear indication for arthroscopy.

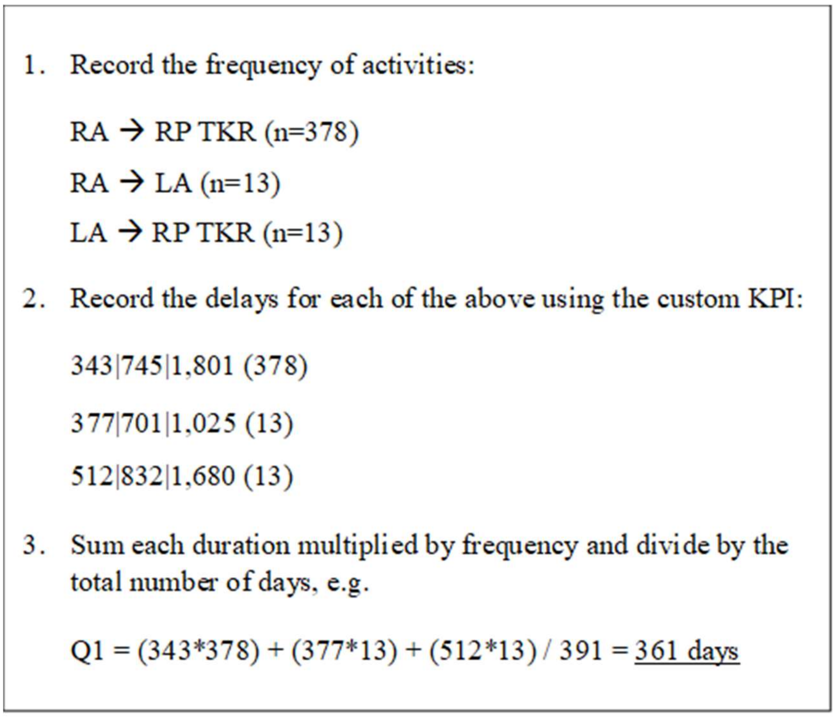
Further analysis was carried out in order to consider the time intervals between surgeries. The custom Connection KPI 2 presented in Section 4.1.3.1 was used to calculate the interquartile range. The screenshot presented in Figure 7.22 shows the Q1, median and Q2 values in days between first right arthroscopy and right primary TKR.

Figure 7.22 Connection between two activities showing the interquartile range for the time interval between surgeries



An example is provided in Figure 7.23 to demonstrate how this information was used to calculate the time interval between two surgeries. As one surgery may directly or indirectly precede another, temporal information between all surgeries along the path must be included.

Figure 7.23 Method for calculation of time intervals between surgeries



1. Record the frequency of activities:

   RA → RP TKR (n=378)

   RA → LA (n=13)

   LA → RP TKR (n=13)

2. Record the delays for each of the above using the custom KPI:

   343|745|1,801 (378)

   377|701|1,025 (13)

   512|832|1,680 (13)

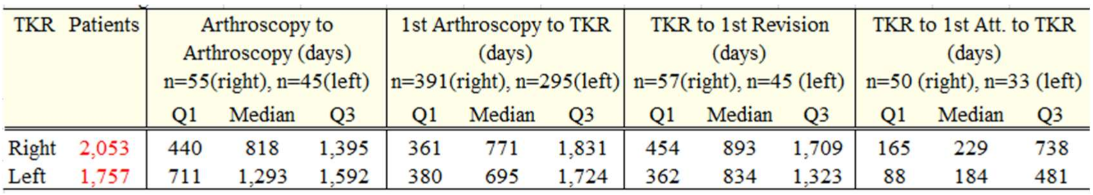
3. Sum each duration multiplied by frequency and divide by the total number of days, e.g.

   Q1 = (343*378) + (377*13) + (512*13) / 391 = 361 days

It must be noted that this is a manual method and when applied to a large number of connections can be extremely time consuming.

The value of 361 days from Figure 7.23 can be seen in Table 7.4 where it represents the lower quartile in days between first right arthroscopy to right primary TKR for patients having TKR surgery on one knee only.

Table 7.4 Time interval between surgeries for SAIL single TKR patients[2]



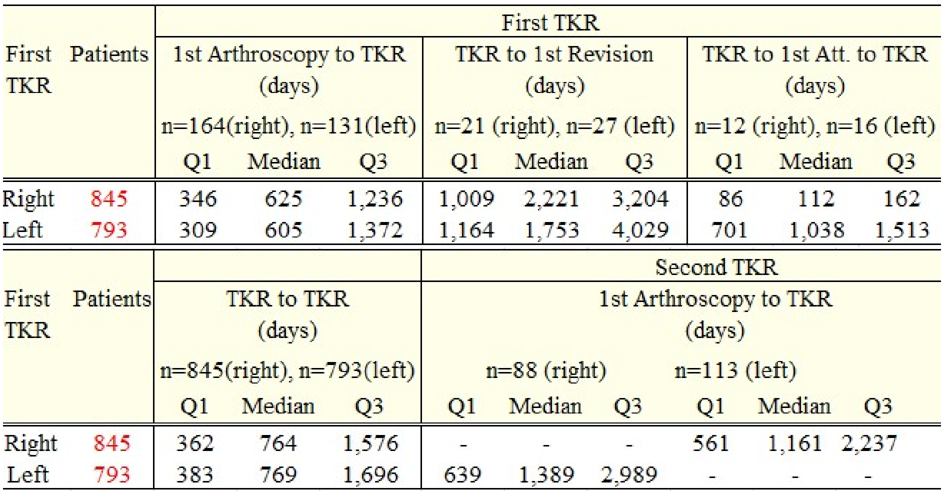
TKR	Patients	Arthroscopy to Arthroscopy (days) n=55(right), n=45(left)			1st Arthroscopy to TKR (days) n=391(right), n=295(left)			TKR to 1st Revision (days) n=57(right), n=45 (left)			TKR to 1st Att. to TKR (days) n=50 (right), n=33 (left)		
		Q1	Median	Q3	Q1	Median	Q3	Q1	Median	Q3	Q1	Median	Q3
Right	2,053	440	818	1,395	361	771	1,831	454	893	1,709	165	229	738
Left	1,757	711	1,293	1,592	380	695	1,724	362	834	1,323	88	184	481

n = patients

As patients may have multiple arthroscopy and revision surgeries on the same knee, to avoid bias, the time intervals between first surgery and primary TKR is calculated. Time between multiple arthroscopy surgeries is calculated as a separate metric. The temporal information presented in Table 7.5 relates to patients who have undergone primary TKR surgery on both knees.

[2] The time interval between primary TKR and first revision includes cases that travel via AttToTKR (right = 9 cases, median time between AttToTKR and revision TKR = 91 days. Left: 7 cases, median time interval between AttToTKR to revision = 427 days.) The time interval between TKR to first AttToTKR includes cases that travel via revision (Right = 4 cases, median time interval between revision and AttToTKR = 125 days. Left: 1 case, time interval between revision and AttToTKR = 216 days.) Time between revisions is not included as the frequency of cases is below 10.

Table 7.5 Time interval between surgeries for SAIL patients with two TKRs[3]



First TKR	Patients	First TKR								
		1st Arthroscopy to TKR (days)			TKR to 1st Revision (days)			TKR to 1st Att. to TKR (days)		
		n=164(right), n=131(left)			n=21 (right), n=27 (left)			n=12 (right), n=16 (left)		
		Q1	Median	Q3	Q1	Median	Q3	Q1	Median	Q3
Right	845	346	625	1,236	1,009	2,221	3,204	86	112	162
Left	793	309	605	1,372	1,164	1,753	4,029	701	1,038	1,513

First TKR	Patients				Second TKR					
		TKR to TKR (days)			1st Arthroscopy to TKR (days)					
		n=845(right), n=793(left)			n=88 (right)			n=113 (left)		
		Q1	Median	Q3	Q1	Median	Q3	Q1	Median	Q3
Right	845	362	764	1,576	-	-	-	561	1,161	2,237
Left	793	383	769	1,696	639	1,389	2,989	-	-	-

n = patients

Important clinical questions can quickly be answered using process discovery. During the final evaluation, two questions were posed by the orthopaedic surgeon. These questions and answers are presented below.

Q1: What percentage of patients undergo "unnecessary" arthroscopies? 5.2%

This is broken down into the following questions which can quickly be answered within Celonis by filtering the event log data, before selecting the number of cases with a time interval of less than a year:

- *What percentage of knee replacement patients have undergone primary TKR surgery with a previous arthroscopy on the same knee?* ***17%***

[3] The time interval between primary TKR and first revision includes cases that travel via AttToTKR (Right = 0 cases. Left: 3 cases. The time interval between TKR to first AttToTKR includes cases that travel via revision (Right = 0 cases. Left = 2 cases.) Time between arthroscopies is not included as the frequency of cases is below 10.

- *Out of those patients that had an arthroscopy, how many had a primary TKR within 12 months of their last arthroscopy?* ***5.2%***

Q2: Does arthroscopy surgery before primary TKR surgery increase the risk of complications and further surgery? Yes

This is broken down into the following questions which can quickly be answered by filtering the event log data within Celonis:

- What is the rate for patients having an arthroscopy followed by a primary TKR followed by a revision TKR, compared to the rate for patients not first having an arthroscopy? **0.046 compared to 0.020**
- What is the rate for patients having an arthroscopy followed by a primary TKR followed by attention to the TKR, compared to the rate for patients not first having an arthroscopy? **0.030 compared to 0.015**

7.6 Stage 4: Evaluation

The aim of this study was to determine whether process mining techniques could be applied to the SAIL data in order to examine surgery pathways for patients diagnosed with knee pain. During this study, a knee pain surgery reference model was defined using process mining techniques and domain expert knowledge. This model was used to identify data quality issues within the SAIL data and to measure the conformance of the data. Finally, the cleansed data was used to generate some episode statistics. Process mining metrics were produced to provide technical assurance to the validity of the models. Construct validity was performed by referencing published scientific literature, where different methods had been used to arrive at similar results. Finally, face validity and clinical plausibility of the results was provided using expert opinion.

7.6.1 Technical evaluation

Verification of the reference model was performed by carrying out semantic analysis on the Petri net model. The model proved to be sound in all areas, as evident in Figure 7.11. This model was used in Section 7.5.1.2 to perform conformance checking where the event log data was replayed over the model to identify data cleansing issues and to assess the quality of the model. For the initial iteration an average replay fitness score was calculated of 98.5 percent. A second iteration was performed in order to cleanse the data in preparation for the next stage where process analytics, in the form of episode statistics, were calculated. After this iteration the average fitness was recalculated at 100 percent. This included ten cases with missing primary events over the entire event log, however, this was caused by a rare variation permitted in medical practice.

7.6.2 Comparison of results

In addition to the verification of the process model, where possible, the results were supported by evidence from within the current literature.

The frequencies of patients undergoing single and subsequent knee replacement surgery were directly compared to those from a similar study using Swedish registry data. All results were within two percent, which helped to provide confidence in the method.

Evidence was found within the literature in support of the finding that fewer knee arthroscopies are performed prior to a second knee replacement [393]. This may be because the patient is younger at the time of the first joint replacement and therefore may have less severe osteoarthritis, warranting an arthroscopy. In addition, a change in the national guidelines [392] to only allow surgeons to perform arthroscopies on patients under strict conditions may have contributed to this finding.

Figures published in the NJR [174] supported the result of approximately two years for the average duration between TKR and first revision surgery.

Our study showed that 5.2 percent of patients undergo potentially unnecessary knee arthroscopies. Werner et al. [393] in their study into TKR surgery after knee arthroscopy, added confidence to this finding by reporting that between 2.2 and 10.2

percent of patients with osteoarthritis who have an arthroscopy will undergo TKR surgery within one year of their arthroscopy.

Finally, the finding that arthroscopy surgery before primary TKR surgery may increase the risk of complications and further surgery was supported by two articles. Liu et al. [394] in a systematic review considering the influence of prior arthroscopy on outcomes of primary total lower extremity arthroplasty and Gu et al. [395] in a study into the association between prior knee arthroscopy and risk of revision, both stated that patients receiving knee arthroscopy before total knee replacement surgery are at a substantially increased risk of revision.

7.6.3 Clinical evaluation

The MSK experts guided the development of study-specific research questions, directed the selection of clinical codes and advised on the reference model rules, before formally evaluating the knee pain surgery reference model and results. The results were first evaluated by an orthopaedic surgeon from LTHT during a structured walkthrough using the interactive knee pain surgery reference model. The orthopaedic surgeon was then joined by a rheumatologist from LTHT and a medical epidemiologist from Union Chimique Belge Biopharma, to perform a final evaluation of the results during a qualitative structured interview. Verification of the work has been continuously carried out during each stage of the study. The purpose of the clinical evaluation is to gather insights into the results, identify potential gaps and limitations in the research and to evaluate the clinical plausibility of the results from a surgical perspective.

During the structured walkthrough, a detailed demonstration was given where the SAIL knee pain data was visualised using the interactive knee pain surgery reference model. The orthopaedic surgeon confirmed that the process mining techniques described in this chapter provided a fast and effective way of generating useful statistics for use by medical and research professionals working in the field of MSK diseases. Specifically, the speed at which it was possible to explore the data to identify possible correlations between variables and outcomes was impressive. To generate similar visualisations using traditional epidemiological/statistical software such as R or STATA would take a considerable amount of programming time and effort. In addition, the process of conformance checking provided the surgeon with an additional layer of confidence in the quality of the data.

The final evaluation was carried out by performing a qualitative structured interview with the three clinical domain experts. The interview was structured using the GQFI Table described in Section 4.1.4.1.2. This table provided the experts with a clear, summarised view of the results, which allowed for informed evaluation. All evidence collected during the previous two sections was entered into the table prior to the interview. For the final validation, face validity and clinical plausibility of the results was discussed. The table, along with detailed results from the interview are presented in Appendix G. A summary of these results are discussed below.

The individual comments from the domain experts can be seen within the table. In summary, the discovered process model used to establish a set of preliminary rules was considered to have a high level of understandability and was in alignment with what the experts would expect. The model was considered useful as it helped to validate the beliefs of the experts with regard to sequence and frequency of surgery events. They accepted the results from conformance checking and thought them to be plausible after visualising the data using the ProM MPE plugin. When guided by an analyst, the experts found the MPE a fast and effective way in which to explore the data, especially for hypothesis generation. They expected the number of violations, especially as routinely collected data was used and attributed the violations largely to miscoding. However, cleansing the data to achieve 100 percent fitness was thought to have been an inefficient use of time. The figures can be used by clinicians when advising patients on knee joint replacement surgery.

The percentage of patients in the most common variants was considered to be useful information to allow for a better understanding of service utilisation and for the provision of data to triage services for planning purposes. The experts explained theories for why fewer arthroscopies may be performed prior to a second knee replacement, some of which were supported by evidence in the published literature.

The frequency of surgeries for patients with single and bilateral TKR surgeries were all as expected, again some results were supported with evidence from the literature. Three reasons were offered as to why there were fewer arthroscopies prior to a second knee replacement. The duration from TKR to revision for bilateral patients was more than double that of patients having single knee surgery. The experts offered a possible clinical explanation for this. This may also be due to outliers which may have skewed the results in such low numbers.

Being able to identify the percentage of TKR surgeries that occur within 12 months of an arthroscopy is an important finding, as the information can be used to make potential efficiency savings. Similarly, knowing whether there is an increased risk for patients having an arthroscopy before primary TKR surgery is useful information. Both these findings are reasonably well known and evidence was found to support them, though they still warrant consideration by the field of orthopaedic surgery.

The overall consensus from the clinical evaluation demonstrated the method, model and results to be realistic, correct, relevant and useful to the overall research programme and knee pain surgery medical domain.

7.7 Future work

Future work in this area may include the addition of more knee pain surgery types to the reference model. The model could be extended to include other knee pain related event types such as physiotherapy referrals, GP visits, tests (e.g. radiograph, magnetic resonance imaging, computed tomography) and prescriptions. Patients may also be stratified based on the severity of their arthritis.

A second analysis could be performed where the primary TKR date is substituted with an 'intention to treat/referral from surgeon' date. The time intervals between first arthroscopy and TKR would then be recalculated. In the second analysis, the three sets of values in Tables 7.4 and 7.5 would then be replaced, before comparing the data to the original analysis. It could then be determined whether patients who had their primary TKR surgery delayed for longer periods had a higher rate of revision type surgeries. Delays to orthopaedic surgery is common especially during the winter months in the United Kingdom, when beds are prioritised for non-elective patients suffering from flu and chest problems.

There is potential in this area beyond surgery. Care pathways may include other events including procedures, referrals, prescriptions, appointments and imaging results. When care pathways are used to help determine future treatment choices, patient-related factors such as health state and disease severity should also be taken into consideration.

7.8 Summary

To date, this study represents the only published research which has applied process mining techniques to data from the SAIL databank. It has applied techniques from all three types of process mining to the SAIL data in order to answer the three study-specific research questions defined in Section 7.3. To answer question one, a knee pain surgery reference model was defined using process discovery, for patients diagnosed with knee pain. Process discovery techniques were used to provide information on the ‘actual’ sequence and frequency of surgery events in the SAIL data. This information was used in discussions with the clinical experts when defining a set of reference model rules. To answer question two, conformance checking techniques were used to confirm that the behaviour in the real-life SAIL data conformed to the newly created knee pain surgery reference model. To answer question three, conformance checking, enhancement and process discovery techniques were all used in order to cleanse and validate the SAIL data before generating some useful healthcare statistics for patients diagnosed with knee pain. Finally to answer question four, a summarised list of the main strengths and weaknesses of process mining SAIL data is presented. In Section 3.5.1 a number of challenges were identified, specifically for process mining healthcare data. Throughout this chapter specific examples of these challenges and how they have been approached have been explained.

Main strengths of process mining SAIL data:

1. Dotted charts provide a high-level view of the event log data. This is useful for identifying data quality issues, errors in the extraction and transformation logic/code, potential bias and hidden patterns of interest.
2. Dynamic control over the level of detail using slider bars, useful when working with data that has a high degree of variation.
3. Grouping of diagnoses into chapters by the clinical coding systems help to simplify data selection code when creating an event log.
4. Business and technical support and comprehensive documentation for the SAIL data.
5. Curated data resulted in fewer data quality issues.
6. It was possible to check compliance against clinical guidelines and best practice.
7. Trace Fitness view identified different data quality issues.
8. The MPE plugin was useful for viewing bilateral and single-sided events.
9. The process models can be validated using publically available healthcare literature.
10. Possible correlations between variables and outcomes may be identified using process mining tools.

Main weaknesses of process mining SAIL data:

1. Left and right sided surgeries caused unforeseen problems. These included complexity of modelling, pre-processing of data and reduction in cohort size.
2. Reduction in the size of the cohort due to challenges with healthcare data such as study windows, legal and ethical considerations and low data quality.
3. High amount of pre-processing, including advanced programming skill was required to create the event log.
4. Clinical domain expertise was required throughout. Availability is often limited due to increasing pressures within the NHS.
5. Lengthy ethical approval process.

Evaluation took place throughout all stages of this study to ensure the results were accurate and meaningful from a clinical perspective. One of the main clinical findings is that by using process mining techniques, patient outcomes can be effectively compared for TKR patients with and without a previous intervention.

Process scientists working in the healthcare domain should be aware that modelling pathways that include sidedness can be challenging and add an extra layer of complexity. As a result, the construction of a model for conformance testing proved much more difficult than it appeared from the literature. In addition to providing some useful episode statistics, the method provided in this section can be followed by other healthcare researchers when working with EHR data. Using the method will provide confidence in respect to the validity of their data and may provide useful insights into some of the challenges and opportunities faced when process mining MSK surgery pathways. The methodological insights gained from this study are intended to advance knowledge in the field of process and data science.

Chapter 8
Discussion and Conclusions

Chapters 5 to 7 presented the analysis of the three studies underpinning the research in this thesis. These chapters were structured using the method described in Chapter 4. The primary research questions posed in Section 1.3 have been broken down and addressed by each of these studies. In this final chapter, the work is summarised before reflections are made on the methods and the findings from each study. This chapter concludes by first assessing whether the hypothesis and primary research questions have been effectively addressed and answered, before discussing the contributions of this thesis, its impact and considerations for future directions.

8.1 Summary of work

This section presents a summary of the work carried out in each chapter of this thesis. Chapter 1 introduced the topic, defined the hypothesis and identified the primary research questions. This was followed by a study of the research approach to be taken, specifically looking at the data sources and datasets to be used. The hypothesis is that process mining techniques can be used to provide insights that may benefit patients suffering with musculoskeletal conditions.

In Chapter 2 secondary research was undertaken into appropriate literature to provide the healthcare background for this research programme. This focused on healthcare systems, clinical coding standards, different methods used to visualise and understand health data and finally, the four types of MSK conditions investigated within this work.

The next chapter reviewed technologies required to underpin the work carried out within the subsequent research studies. This involved examining database management systems, data and process science, process modelling techniques, approaches to process mining, specifically process mining within the healthcare domain and associated challenges. A structured search was carried out to review process mining within MSK conditions. Different evaluation techniques were discussed before providing a summary of both the healthcare and technical secondary research.

Chapter 4 described the method used for the practical studies. It began by providing an overview of how the PM^2 method had been adapted for use in this work, before describing common aspects of its implementation for each of the five stages for this work. The final part of this chapter expanded upon the introduction to the three datasets described in Chapter 1, to provide a more detailed description of the data.

Chapter 5 presented the first of the three practical studies by applying process mining techniques to the MIMIC-III data in order to create disease trajectory models. This study drew upon a disease trajectory model published by Jensen et al., where a gout diagnosis was found to be central within the cardiovascular cluster of diseases. An iterative approach was taken to this study, where Jensen's rules were gradually applied to the MIMIC-III data by implementation of 10 sets of data transformation rules. An introduction to the work was presented before moving through the first four stages of the method. During the planning stage, the aim was defined and five study-specific research questions were composed to satisfy the primary research question in Chapter 1. The Extract, Transform and Load (ETL) stage first included an overview of the Jensen method, before describing the specific ETL steps necessary to create the event logs required for process mining and analysis. The DISCO process mining tool was used to discover disease trajectory models from eight event logs during Stage 3. Two of these models were presented within the main body of this thesis, the first created by replicating as closely as possible Jensen's rules and the second by making refinements to these rules. In Stage 4 the two models were evaluated, first by comparison with the model created by Jensen, and secondly the findings were discussed and evaluated for correctness by an expert in the field of cardiovascular medicine. No process improvements were made as the work was purely research, therefore Stage 5 of the method was not included. Finally, impact of the work was presented before providing a summary of the chapter.

The second of the practical studies is presented in Chapter 6. This study used data collected via the MyPathway mobile application from MSK patients attending STHT. During this study, the patient data was stratified by health outcome, which was determined by comparing PROM data over time. Process mining techniques were applied to the datasets to identify potential indicators of a good or bad health outcome. Changes were also identified in order to help improve data collection for future process mining studies. Within this chapter, an introduction to the study was presented before six study-specific research questions were outlined during the planning stage of the method. During the ETL stage, an overview of the MyPathway system and MSK care pathways at STHT was provided before describing how health outcomes had been defined using EQ-5D data. Due to the large number of data quality issues present in the data extract, the data transformation stage was separated into data cleansing and data processing. Forty event logs were created for process discovery purposes using the Celonis software. These event logs contained knee and spinal pain patient data and were separated by health outcome, age group and sex. Characterisation of the datasets was performed using R Studio and Excel in order to more accurately interpret the analysis results. Clinical evaluation of these results was carried out by the Clinical Service Manager at STHT. Stage 5 discusses how the developers of the MyPathway software have used this work to improve their processes. Finally, ideas for future work were discussed before providing a summary of the chapter.

Chapter 7 presented the third practical study for this research programme. Here, process discovery, conformance and enhancement techniques were applied to the SAIL dataset to help create an expert-defined interactive reference model for knee pain surgery and to generate some episode statistics. During the planning stage, three study-specific research questions were defined. In Stage 2 the steps included in ETL were described. Dotted charts were presented after they were used to verify the semantics of the extraction code. Data transformation activities involved the aggregation of information to create left and right-sided high-level knee pain surgery events. The majority of the process mining for this study was carried out using the ProM framework. However, in order to understand the general control flow of the knee pain surgery process before meeting with the domain experts an initial process model was discovered using Celonis. After creating an interactive knee pain reference model, using all three types of process mining, the model was used to help cleanse and validate the SAIL data in preparation for loading into Celonis for the generation of episode statistics. Process analytics were presented and evaluated against results from a study using data from a different health dataset, the Swedish Knee Arthroplasty Register. In addition, technical evaluation of the model quality was carried out using process mining quality metrics and clinical evaluation of the findings was performed using the GQFI Table to carry out a structured interview with three clinical domain experts. An iterative process with an expert (orthopaedic surgeon) enabled refinement of the work, enabling understanding of limitations and identification of areas for future research.

8.2 Reflection on the method

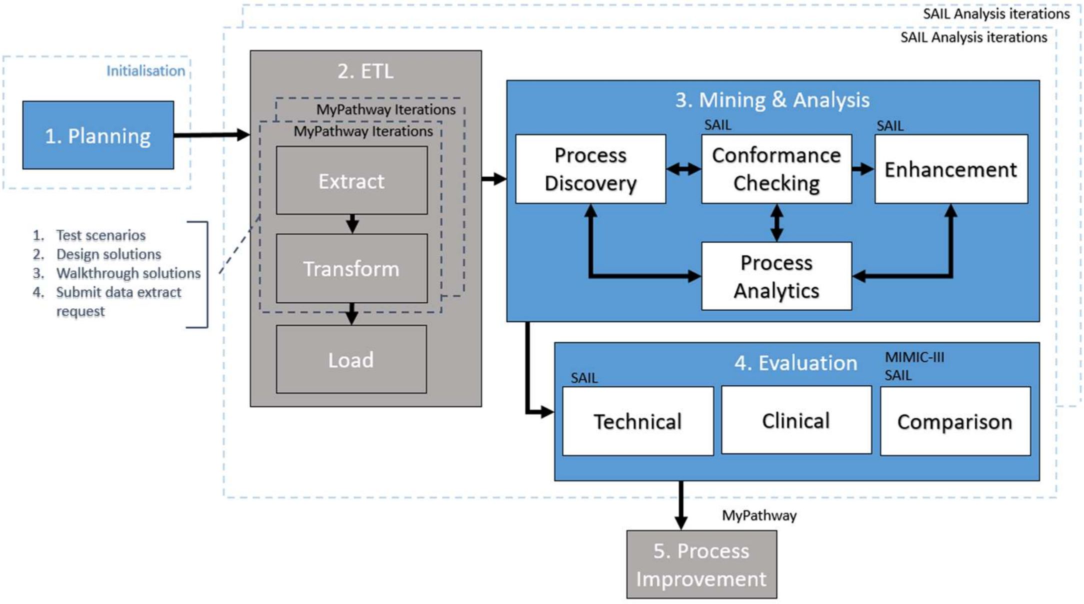
The following hypothesis was stated in Section 1.3: '*routine healthcare data can be analysed using process mining techniques in order to provide insights that may benefit patients suffering with musculoskeletal conditions.* To test this hypothesis three studies were performed and are presented in chapters 5 to 7. Within each study, a variety of process mining techniques were applied to three different routinely collected datasets. This section reflects on the method chosen for these three studies.

The method is amended from PM^2, as presented in Section 4.1. To provide consistency throughout this thesis a decision was made to adopt a single method for all three studies. Finding a suitable method that allowed for a flexible implementation was challenging. The three reasons it was necessary to amend the PM^2 method for use with this research were: 1) the datasets originated from different data providers, each with their own set of data access rules; 2) the quality of the extracted data varied considerably between the datasets and it was necessary with the MyPathway data to perform 25 iterations of data extraction and transformation. For each iteration the data extract was tested against the different business scenarios, new solutions were designed and tested during walkthrough sessions, before submitting a new data extract request; 3) different process mining techniques were explored in the three studies.

When working with the SAIL data dotted charts were used to verify the semantics of the data extraction code, therefore multiple analysis iterations were performed.

Figure 8.1 presents an overview of the implementation of the method applied during the three studies. Stages coloured in blue indicate no change to the PM^2 method, whereas stages coloured in grey signify customisation of the method for this research.

Figure 8.1 Implementation of method based on PM2 [16]



During the Mining and Analysis stage, process analytics and discovery activities were performed in all studies. However, only in the SAIL study was conformance checking and enhancement activities carried out. Clinical evaluation of the results was performed for all studies. In addition to clinical evaluation, technical evaluation was carried out for the SAIL results, also the MIMIC-III and SAIL results were evaluated by comparing them with similar peer reviewed studies. As is evident in chapters 5, 6 and 7, verification occurred throughout the entire process at each stage, where output was assessed by the domain experts prior to the next step. Finally, it was only appropriate to carry out process improvement activities for the MyPathway study.

The documented steps of the PM^2 method helped to guide the research through each stage of development by identifying the inputs and outputs at each stage, along with a set of recommended activities and artefacts. Using this method helped to provide the author with a framework for rigorous, relevant and reproducible research and the reader with a standard format for all three studies. However, evaluation is a critical part of any research project and the method would benefit from more guidance during this stage. More detailed guidance on the different verification and validation techniques available for process models and their related results and findings may help practitioners. This reflection is backed by a recent review of the literature [317] where process mining projects were found to lack a systematic approach for determining the accuracy and meaning of their findings.

8.3 Discussion of the three studies

The previous two sections have summarised the work in this thesis and reflected on the method used to carry out the three practical studies for this research programme. In this section, the positive aspects of each study, the limitations of using each of the datasets and the conclusions for each of the studies are discussed.

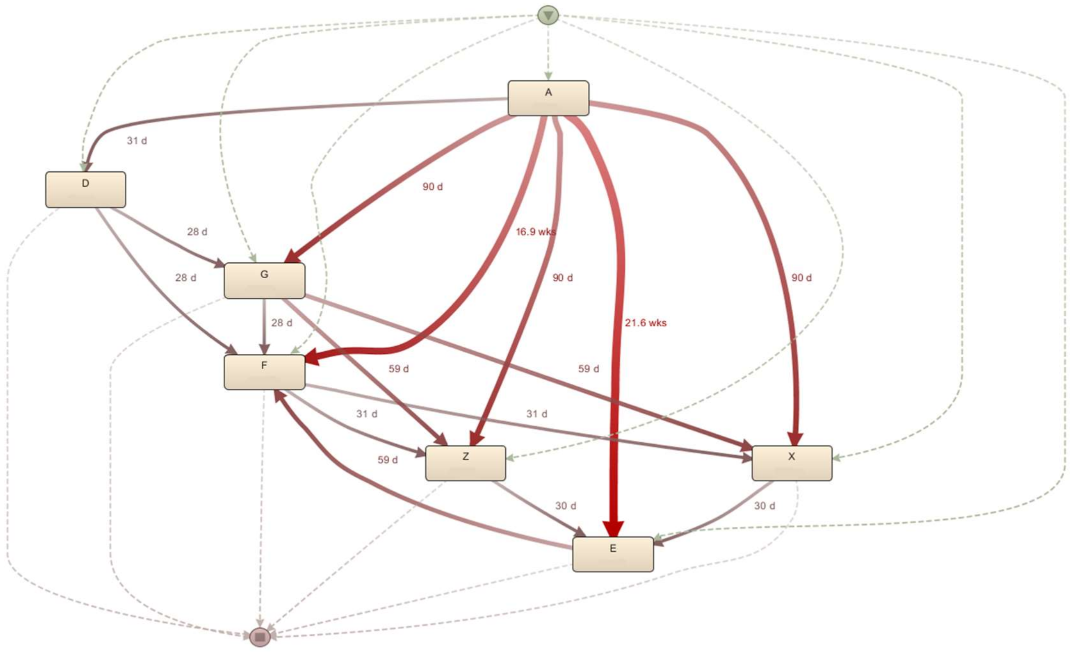
8.3.1 Discussion of MIMIC-III study

The positive and negative aspects along with the conclusions for the MIMIC-III study will now be discussed.

8.3.1.1 Positive aspects of the MIMIC-III study

This study reproduced the CVD trajectory model reported by Jensen, using process mining and the MIMIC-III data. The results of the study showed how process mining techniques could be used to produce disease trajectory models from event log data. However, within the study, time durations between directional ordered pairs of diagnoses were not discussed as time was not included on the Jensen model. It is important to note however, that using process mining techniques, the time intervals between diseases is easily calculated, an example of this is provided in Figure 8.2.

Figure 8.2 Disease trajectory model displaying the median time intervals between diseases



The disease trajectory model presented above has been created using the DISCO process mining tool and the synthetic data described in Section 5.3.3. Here, the median frequency of each directional ordered pair is displayed. Mean, maximum and minimum time intervals can also be displayed.

This work has opened up future opportunities for a wide range of multi-disciplinary research with the potential of providing new insights that may benefit patients suffering with musculoskeletal conditions . A co-authored literature review on the use of process mining techniques for the creation of disease trajectory models [362] showed there to be only four publications in this area. One of which was co-authored [361], another cited the author [329] and the remaining two were published following the presentation of a poster at an international conference [360].

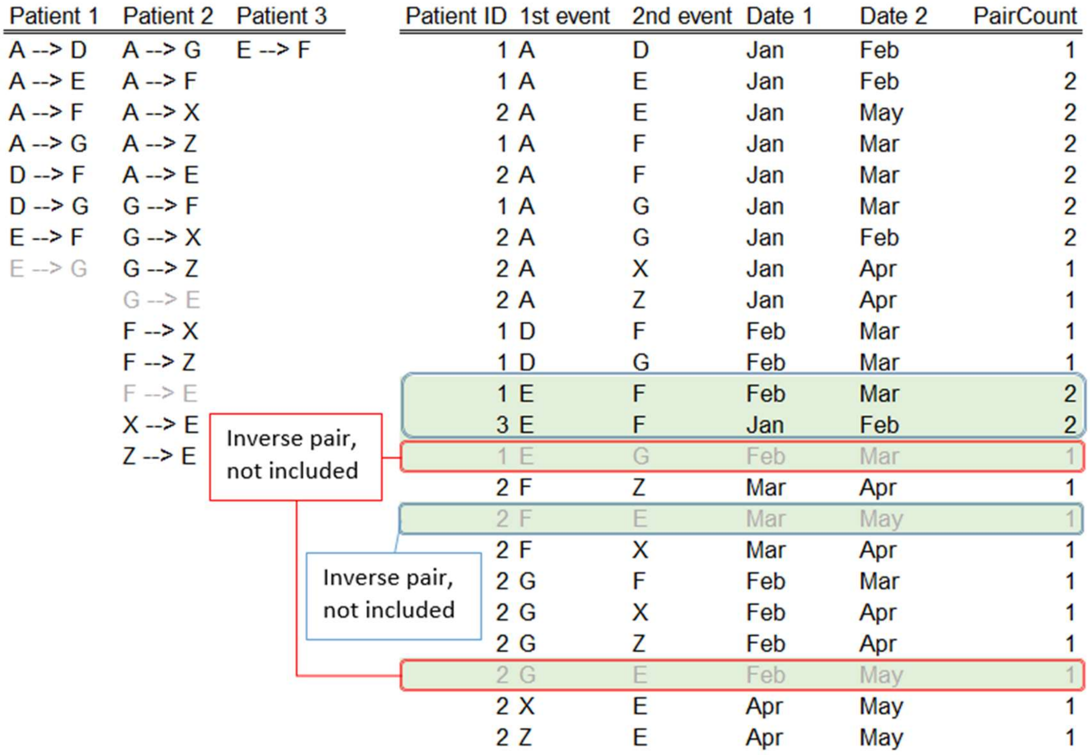
PROMs were used to create separate event logs based on the patient’s perceived level of improvement. Using the data in this way provided an effective method for filtering the event logs in order to make comparisons between different cohorts of patients.

8.3.1.2 Limitations of the MIMIC-III study

Due to the complexity of the data transformation rules for the two experiments (see Figure 5.7) an experienced programmer was required to assist with the pre-processing of the event log data.

The second limitation is concerned with the loss of patient information in the disease trajectories. When using directional ordered pairs to create disease trajectory models with process mining techniques it is not possible to retain the patient information. The reasons for this are explained using the data from Figure 8.3.

Figure 8.3 An example to illustrate processing of event log data



Patient 1	Patient 2	Patient 3
A --> D	A --> G	E --> F
A --> E	A --> F	
A --> F	A --> X	
A --> G	A --> Z	
D --> F	A --> E	
D --> G	G --> F	
E --> F	G --> X	
E --> G	G --> Z	
	G --> E	
	F --> X	
	F --> Z	
	F --> E	
	X --> E	
	Z --> E	

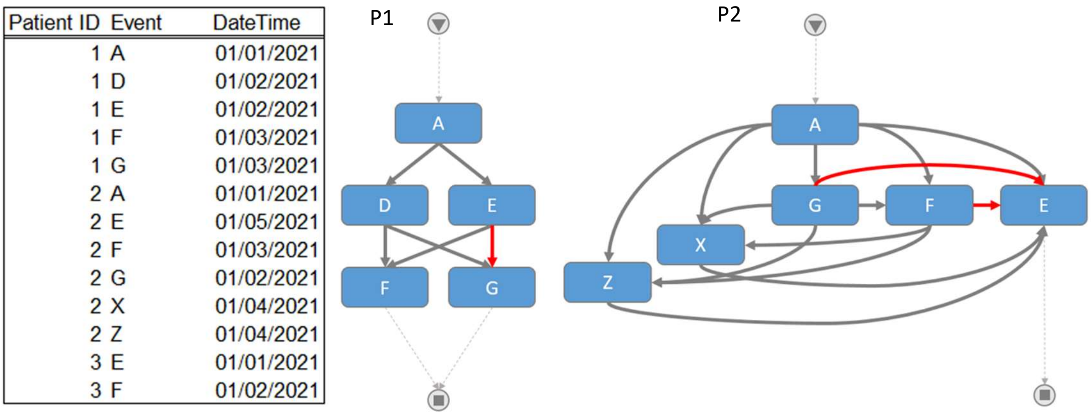
Patient ID	1st event	2nd event	Date 1	Date 2	PairCount
1	A	D	Jan	Feb	1
1	A	E	Jan	Feb	2
2	A	E	Jan	May	2
1	A	F	Jan	Mar	2
2	A	F	Jan	Mar	2
1	A	G	Jan	Mar	2
2	A	G	Jan	Feb	2
2	A	X	Jan	Apr	1
2	A	Z	Jan	Apr	1
1	D	F	Feb	Mar	1
1	D	G	Feb	Mar	1
1	E	F	Feb	Mar	2
3	E	F	Jan	Feb	2
1	E	G	Feb	Mar	1
2	F	Z	Mar	Apr	1
2	F	E	Mar	May	1
2	F	X	Mar	Apr	1
2	G	F	Feb	Mar	1
2	G	X	Feb	Apr	1
2	G	Z	Feb	Apr	1
2	G	E	Feb	May	1
2	X	E	Apr	May	1
2	Z	E	Apr	May	1

Inverse pair, not included

Inverse pair, not included

The information above again uses the same synthetic data presented in Section 5.3.3. On the left, the ordered directional pairs for each of the three patients are listed, with ones for deletion shown in grey. The table to the right explains the processing behind these deletions. Using the same threshold value of 40/60, two rows are highlighted in red for deletion (E→G and G→E) as there is one instance of each and therefore neither direction reaches this threshold value. For the three rows highlighted in blue only the pair F→E for patient 2 will be deleted as the inverse (E→F) is has two instances. When all repeating diagnoses for patients are removed, using a directly follows graph, when creating the event log using the patient as the Case ID the process models presented in Figure 8.4 should be produced.

Figure 8.4 Process models for patients 1 and 2



Patient ID	Event	DateTime
1	A	01/01/2021
1	D	01/02/2021
1	E	01/02/2021
1	F	01/03/2021
1	G	01/03/2021
2	A	01/01/2021
2	E	01/05/2021
2	F	01/03/2021
2	G	01/02/2021
2	X	01/04/2021
2	Z	01/04/2021
3	E	01/01/2021
3	F	01/02/2021

The event log in Figure 8.4 represents the data in Figure 8.3 with all repeating patient diagnoses removed. Using this event log data, it is not possible to create the two process models for patients 1 and 2 (P1, P2), as the two connections, highlighted in red, will always be incorrectly generated. To overcome this problem, the case ID must refer to a directed ordered pair, resulting in the patient ID being lost.

The third limitation relates to the method described in Figure 5.3. Creating a window of interest, therefore deleting all secondary diagnoses for the first admission, assumes that all existing comorbidities for a patient are recorded at the first admission, which is not always the case.

The absence of any form of hold back or cross-validation method (see Section 3.6.3) is the fourth limitation for this study. These methods would provide more confidence to the results. However, the work in this chapter laid the foundations for the work on disease trajectory modelling by Kusuma, which led to two joint publications [361], [362]. K-fold cross-validation was identified as an appropriate technique and was subsequently implemented by Kusuma on disease trajectory models using MIMIC-III data [329].

The remaining limitations discussed in this section relate to the two datasets used to create the disease trajectory models. As previously discussed, gout is primarily diagnosed and managed by GPs and is therefore rarely recorded in secondary care systems. Both studies used data collected from secondary care systems, albeit, for the Jensen study the data represented a larger percentage of the population. Had the data originated from primary care systems, it is possible that the link between gout and cardiovascular diseases would have been stronger.

The fifth limitation is concerned with the purpose behind the recording of diagnoses, as this may affect the results. As mentioned in Chapter 5, coded MIMIC-III data is used primarily for billing and administrative purposes rather than quality of care, as it originates from a private hospital in the United States. Whereas, the majority of hospitals submitting data to the Danish National Patient Registry, used in the Jensen study were publically funded. As a result, it is possible that bias may be introduced to the results in response to certain diagnoses yielding a higher revenue and others going unrecorded.

The final limitation is also related to the differences between the two datasets. There will undoubtedly be disparities between the models due to the differences in patient behaviour caused by financial implications. All Danish residents receive publicly funded primary and secondary health care, which is largely free at the point of use, whereas patients attending the BIDMC in Boston pay privately for their care.

8.3.1.3 Conclusions of the MIMIC-III study

Drawing on the positive and negative aspects of the study, it can be concluded that process mining techniques can be effectively used to create disease trajectory models. Using event log data and standard software tools, process mining techniques can be used to automatically draw disease trajectory models. This is a major step forward when compared to the hand-drawn diagrams created by Jensen [38]. By extending this work [329] to provide a pipeline in order to standardise much of the pre-processing programming effort, disease trajectory modelling will be available to a wider range of people.

A process-mining approach also allows for the temporal information between diseases to be preserved. In addition to the automatic creation of the diagram, standard tools can also be used to calculate the mean, median, minimum and maximum time intervals between diseases. The inclusion of temporal data is a significant improvement from the models produced by Jensen, as the time intervals between diagnoses can be observed. However, when producing disease trajectory models using directed ordered pairs, as with the method described in Chapter 5, individual patient journeys cannot be traced.

One of the most valuable pieces of information prior to using any method, would be to know whether a diagnosis was pre-existing for a patient. As this information is not available, before the data transformation code can be designed consideration must be given to how the diseases are recorded in the source systems. For example with the MIMIC-III data, because the majority of cardiovascular diseases are chronic in nature, a window of interest was created and repeating diagnoses for patients were discarded. A less severe approach may be to allocate a threshold value to specific disease types. Here, different diseases would only be discarded if they were re-recorded in the patient data before that amount of time had elapsed. In all cases, it must be remembered that these type of design decisions must always be guided by expertise from the clinical domain experts. The removal of repeating patient diagnoses increases the sensitivity of the model which was apparent with the second model as it lead to the discovery of more acute diseases which were previously hidden by large amounts of repeating diagnoses.

Reflections on the use of the MIMIC-III dataset have found it appropriate for use in the reproduction of the model created by Jensen, as both datasets were constructed from secondary care data and therefore generally comparable. Using Jensen’s data extraction and transformation rules made it possible to validate the method by comparing the two models. The Jensen model showed gout to be a central disease with the CVD cluster. However, our results did not show such a strong association and required the model to be adjusted to include the top 12 percent of diseases. The association is likely to be caused by a combination of its inflammatory component and the use of certain medications for CVD. As gout is most often diagnosed and managed by GPs, using linked primary and secondary care data may provide a more accurate picture when exploring its existence within the cardiovascular cluster of diseases.

8.3.2 Discussion of MyPathway study

The positive and negative aspects along with the conclusions for the MyPathway study will now be discussed.

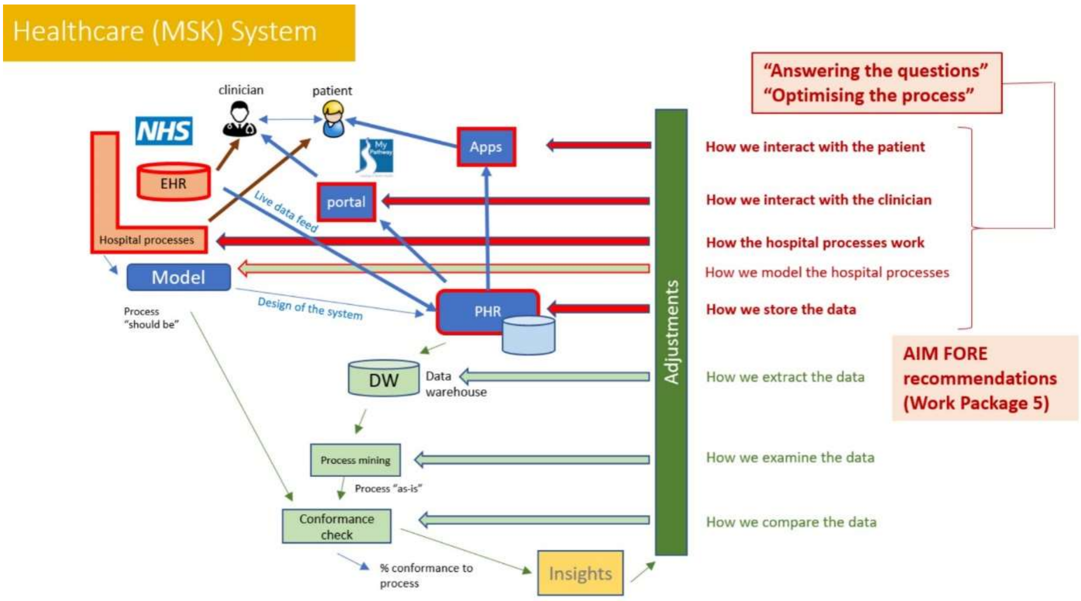
8.3.2.1 Positive aspects of the MyPathway study

An area seldom discussed within the literature is the activities involved with event log generation [396], despite the fact that they often account for the majority of the time and effort in process mining projects [397]. A large proportion of the challenges faced during this study were in connection to the generation of the event log. The work presented in Section 6.3 has helped to fill this gap by describing in detail the various steps involved in this process.

Working as part of the MyPathway team provided the opportunity to explore the challenges and opportunities of working with uncurated healthcare data. In addition to the data, this also provided direct access to the clinical, business and technical experts of the MyPathway system, something that is often not possible when working with other datasets, particularly research datasets.

The work in this study has contributed towards many changes that have been implemented by ADI and STHT in order to improve their processes. The process mapping diagrams were used in presentations to inform clinicians and other stakeholders. Many of the inference and data cleansing rules have been incorporated into the tool by the development team at ADI to assist with the shortcomings in the data acquisition, detailed in Section 6.3.5. Amendments to business processes have been made, such as a change to when certain questionnaires are allocated to patients. The diagram presented in Figure 8.5 was included in the AIM-FORE final report to describe the high-level architecture for the future direction of the MyPathway system.

Figure 8.5 Future MSK system (reproduced with permission from ADI)



This diagram shows that the clinician interacts with the patient via the application and the clinician accesses the application via the portal. There is a live data feed from the hospital systems with the patient's EHR. Data from the EHR is extracted, transformed and loaded as a Patient Health Record (PHR) into the MyPathway system where it is used within the application. The ideal or 'should be' hospital processes are modelled and used to ensure that the functionality of the system meets the requirements for the hospital. To provide insight into patient behaviour, process mining techniques are used. For this, key data is copied to a data warehouse and process discovery techniques are applied to the data in order to create process models where the 'actual' behaviour of patients can be analysed. Conformance checking techniques can be applied by comparing the 'should be' models against the 'as is' models to help identify any deviations. Insights are then fed back and adjustments are made if necessary.

The work undertaken for the MyPathway study helped to create this high-level architecture. To allow for an effective and sustainable system, the questions on the right side of the diagram must be answered and the responses kept up to date with changing requirements and technology.

8.3.2.2 Limitations of the MyPathway study

The first limitation to this study was the amount of time and effort required during the data extraction and transformation stages due to the high number of data quality issues. There were three main reasons for this: 1) limited technical resources for data extraction from the source system; 2) implementation issues surrounding NoSQL type databases; and 3) the MyPathway system was still in its infancy. Twenty-five iterations of data extraction and transformation were performed over an eight month period, consuming much of the total research time. From the many data quality issues listed in Chapter 6, the main ones included: missing relationships for patient referrals; missing event attributes, for example non-labelling of questionnaire and appointment types; incorrect timestamps, for example all inpatient discharges were timestamped with the admission time, many appointments were recorded retrospectively; and imprecise timestamps that had been generated by batch programs. Due to these data quality issues, approximately 36,000 patient referrals were unable to be fixed and were therefore excluded from the data processing stage. This significantly reduced the number of potential patient referrals for analysis and rendered certain candidate event types unusable. In addition to the impact on the amount of usable data, it was necessary to enlist an experienced programmer due to the complexity of the pre-processing rules.

The second limitation relates to the absence of data items that may have proven useful when identifying indicators of a health outcome. Three potentially useful pieces of information have been identified, these are: 1) to know when a patient had read a resource, therefore indicating active use of the application; 2) sensory information such as the patients' step count or heart rate. Functionality existed in the MyPathway software to record this kind of information, however, the feature had not been implemented at STHT; and 3) the patient diagnosis. Had this been present in the data, it could have been used to stratify the patient referrals, rather than using the part of the body for which the patient was referred.

The third limitation is that only patients registered to use the MyPathway application received questionnaires. Therefore non-MyPathway patients could not be assigned a health outcome. Due to this, it was not possible to draw conclusions as to whether the use of MyPathway application could be a possible predictor of a health outcome.

8.3.2.3 Conclusions of the MyPathway study

Initial analysis of the results showed three possible indicators of a poor health outcome, these were: 1) time between GP referral and first appointment. On average, patients with a declined health outcome attended their first appointment six days later than those who improved for both knee and spinal pain patients; 2) patients attending phone appointments, especially older spinal pain patients and male knee pain patients; and 3) patients that did not attend their appointment (DNA), especially younger knee pain patients. Over twice as many young knee pain patients with a declined health outcome did not attend their outpatient appointment, compared to those with an improved health outcome. However, after clinical evaluation of the results, the expert opinion was, that due to the small margins of difference in combination with the relatively small numbers of patients, these three indicators may be down to chance. It can therefore be concluded that process mining techniques cannot be applied to the MyPathway data in order to identify possible indicators of a patient health outcome. It must be noted that this limitation is due to the dataset, rather than the methodological approach.

Although the MyPathway data has proven to be unsuitable for use when applying process mining techniques in order to identify indicators of a health outcome, many lessons have been learnt. Valuable insights have been leveraged from this study and discussed to assist future researchers. The work carried out and presented in Chapter 6 has had impact regarding the improved efficiency of the current MyPathway software and the future direction, with regards to process optimisation for the mobile application platform ADI.

NoSQL databases such as MongoDB are gaining popularity but, as they become more common, process mining and other applications where data is needed to be extracted may become more difficult. This type of database system is ideal for providing fast retrieval of health-based text and other unstructured data often found in EHRs [398]

(Section 3.1.2) though unless specifically set-up, reporting is difficult. Software providers of these database management systems (DBMS) often make provision for the transfer of an organisations' data from their traditional relational DBMSs [399] into the new system. Though consideration is rarely given to the transfer of data in the opposite direction. In these cases, as experienced with the MyPathway data, the extracted data will suffer from low referential integrity. A limited amount of linkages will exist in the data, causing issues when the data analyst needs to relate case events and attributes. The extraction of the MyPathway data should have taken no more than a week, however due to the difficulties caused by the NoSQL structure of the extracted data it took 12 months. In future these problems can be avoided by software developers creating event logs where rules are defined to create mappings between the case events and attributes. These rules should then be embedded into the database systems from the start so as they populate event logs automatically as the system is in operation. Process mining may then be performed using that event log.

Models discovered using process mining techniques can be useful, though when information from multiple models is being compared careful consideration must be given to how the results are presented. In order to report the results in a meaningful way, data may need to be consolidated and presented using a selection of tables and charts. Thought must be given to how the results will be evaluated. Due to the number of potential false positives, when analysing multiple variables, all with equal importance, it may not be possible to apply statistical significance testing to determine whether the differences between the results are reproducible. This may be difficult when performing exploratory studies, as it is not possible to know key variables from the start of the project.

8.3.3 Discussion of SAIL study

The positive and negative aspects along with the conclusions for the SAIL study will now be discussed.

8.3.3.1 Positive aspects of the SAIL study

The aim of this study was to determine whether process mining techniques could be applied to the SAIL data in order to explore surgery pathways for patients with knee pain. The work presented in Chapter 7 of this thesis has successfully fulfilled this aim.

Visualising the data at a high-level using the dotted chart in ProM helped to verify the data extraction rules and code. By using these charts it was possible to quickly sort the data in multiple ways to spot systemic bias and potential errors at an early stage. Using dotted charts at the beginning of a study is also a good way to generate hypothesis within exploratory research projects. During the early stages of the study using process discovery techniques proved effective when helping to establish a set of preliminary rules when designing a knee pain surgery reference model. Later, conformance checking, enhancement techniques and various plugins within the ProM framework, together with related technologies were explored to arrive at an optimal method that could be used to identify data cleansing issues, provide quality metrics and identify potential model refinements. Data discovery techniques were applied to the cleansed and validated data to demonstrate how easily important clinical questions could be answered and episode statistics could be generated using process mining techniques. After evaluating the work an orthopaedic surgeon and two MSK experts found that process mining techniques provided a fast and effective way of generating useful statistics and answering clinical research questions. Specifically, the speed at which it was possible to explore the data in order to identify possible correlations between variables and outcomes was found to be impressive, especially when compared to more traditional techniques.

The modelling and analysis of bilateral events is common within a healthcare setting, as people have a right and left side. However, the challenges associated with sidedness are seldom discussed. During this study, many of these challenges were experienced when working with left and right-sided surgeries. These challenges were discussed and solutions provided.

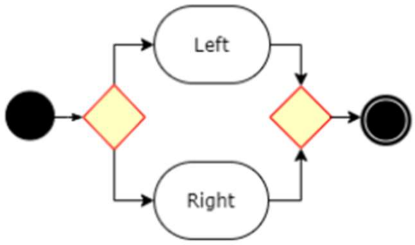
Finally, Erdogan and Tarhan performed a systematic mapping of process mining studies in healthcare [281]. During this study they noted a lack of studies covering cases from more than one hospital. The SAIL study has helped to fill this gap by applying process mining techniques to healthcare data gathered from over 40 NHS hospitals across the entire country of Wales [400].

8.3.3.2 Limitations of the SAIL study

For the purpose of testing the feasibility of the model and the method, only primary terms were used to identify the cohort of patients. When reproducing this method to investigate trends into knee pain surgery a sensitivity analysis should be performed. During the sensitivity analysis, patients should be selected using both primary and secondary knee OA and knee pain terms.

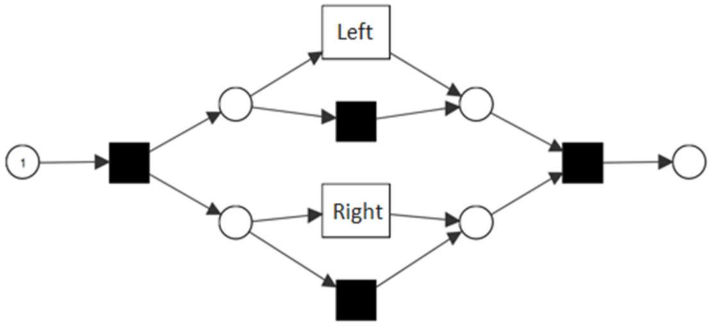
The inclusive OR (OR) construct is often required to accurately model healthcare processes, as it is often impossible to know in advance whether a patient will require left and right-sided care. For example, in this study a patient may have any combination of left and/or right-sided surgeries, but they must have at least one: <Left> or <Right> or <Left→Right> or <Right→Left>. This is modelled in Figure 8.6 using a UML activity diagram.

Figure 8.6 Inclusive OR modelled using a UML activity diagram



However, when conformance checking in ProM a Petri net model is required. This presents a problem when needing to model free choice, as they only allow for AND or exclusive OR (XOR) logic [401], therefore compromises need to be made. For this study a decision was made to allow for cases with no behaviour in the model. A simplified version of the construct is presented in Figure 8.7.

Figure 8.7 Modelling free choice using Petri net notation



This meant that during conformance checking, any patients without left or right-sided surgery would not be flagged as a violation. In this instance it was not an issue, as patients with no behaviour would not be included in the event log.

As mentioned above, a Petri net is needed in order to perform conformance checking in ProM. Despite there being standards for exporting a schema from drawing tools, for example BPMN models in Cawemo [402] or UML Activity Diagrams in diagrams.net [403] there is no facility within ProM to accept an XML file as input for conformance checking. This limitation presented a challenge during this study.

All other limitations identified during this study relate to the data. The first is that there is potential bias in the data and the level of representation which is caused by various unknowns. The results presented in Section 7.5.2 are on the premise that the patient needed the surgery at that point in time. There are many reasons why patients may not have received surgery, some of these may include: knee replacement surgery was recommended by an orthopaedic surgeon, but the patient refused; the patient requested knee replacement surgery, but the orthopaedic surgeon refused as they thought it is too soon; both surgeon and patient agreed that knee replacement surgery was the best option, however the patient was too unfit; the patient was unwilling to undergo surgery on the second knee, as they had a bad experience with the first; and different surgeons often have different judgement thresholds for if and when surgery should be performed.

Other limitations due to the data relate to the duration of a minimum of five years' worth of follow-up time after the date of the patients' first knee pain diagnosis. In addition, the need to simplify the reference model in order to make it usable required the restriction of some less common surgeries.

Pain is subjective and there is a definite discordance between patients that have knee pain and those that have knee OA [404]. However, it is widely recognised that osteoarthritis, including knee OA, is under-recorded within primary care in the UK [405] and therefore to ensure most cases were captured, knee pain patients were included in the dataset.

8.3.3.3 Conclusions of the SAIL study

After drawing on the positive and negative aspects of the study, it can be concluded that process mining techniques can be effectively used to examine surgery pathways for patients in the SAIL data with a diagnosis of knee pain including knee OA.

The SAIL databank is an excellent resource for researchers requiring high quality anonymised, longitudinal, primary and secondary care linked data. Data within SAIL covers the entire population of Wales. The data access process is simple, fast and comparatively inexpensive, especially when compared to similar resources such as ResearchOne [406] and Clinical Practice Research Datalink (CPRD) [407] linked to Hospital Episode Statistics (HES) [408]. Access is granted via a two stage application process that includes Information Governance Review Panel (IGRP) approval. This process is usually completed within 12 weeks. Remote access to the data has proven invaluable during the Covid-19 pandemic with a continued high level of support provided by the SAIL analysts and technical staff.

Both primary and secondary care patient EHRs are required when performing research into total knee replacement (TKR) surgery pathways using historical data from patients diagnosed with knee OA. When using the SAIL data, it was necessary to select the diagnosis data from the GP dataset and the surgical data from the hospital dataset, as the quality of hospital data in the GP dataset was unreliable. For example, the GP dataset contained records for patients with multiple primary TKR surgeries on the same knee and often the laterality information was missing. When this data was checked against the hospital datasets, the dates for the procedures recoded in the GP datasets also proved unreliable. A previous study [152] used non-linked CPRD data to estimate the lifetime risk of undergoing TKR surgery. After speaking with the author, it was clear that similar issues occurred in their study. Within the CTV-2 data, laterality was not recorded. Therefore, where more than two primary TKR surgeries appeared for a patient they selected the first occurring primary TKR Read code within the timeline.

No similar published studies currently exist where process mining techniques have been used to analyse knee pain surgery data. A number of different process mining related techniques have been applied and discussed during this study, particularly in the area of conformance checking. A knee pain surgery reference model was created and used to check the conformance of the SAIL data. After using conformance checking to help identify data cleansing issues, process discovery techniques were applied to a cohort of patients from the SAIL data in order to generate some statistics related to TKR surgery. These results closely aligned to those from similar study that used non-process mining techniques and were found to be realistic and useful when evaluated by clinical domain experts. Finally, the clinical experts found the process mining techniques used within this study to be effective when answering import clinical questions from patient EHR data.

8.4 General discussion

The studies take three different perspectives that demonstrate the breadth of process mining support for healthcare processes within MSK. Many of the techniques used could have been applied across all studies, however the purpose was to explore as many different techniques as possible within the time. Results generated from using any method are dependent on data inputs and process mining is no exception. The technical background chapter (Section 3.5.1) discussed some of the challenges specific to process mining in healthcare. These three studies have confirmed and suggested solutions to some of these challenges as well as identified additional challenges.

Event log generation is a critical and often time consuming part of a process mining project. Different considerations needed to be taken into account with all three datasets and during the studies detailed steps and techniques were suggested. There are many factors to consider when extracting data for event logs.

The MyPathway study was hampered by particularly challenging data quality issues mainly because of two reasons. Firstly, our study was the first to carry out data analysis on the uncurated data and secondly, the technology used to store the data was a NoSQL database (Section 3.1.2) where no thought had been given to data extraction for non-operational purposes. Throughout the studies different techniques were used to manage data quality issues. The Mans Data Quality Matrix provided a framework for analysis of data quality issues. Conformance checking provided a means to investigate the data and generated metrics. The dotted chart presented a way in which to explore the data at various levels of abstraction in order to help establish the root cause of data quality issues. Using trace clustering techniques, patterns of interest were identified from a small number of cases that may of otherwise been missed.

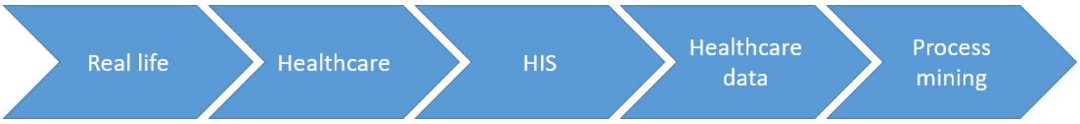
Creation of the data exclusion criteria for the event log is an important step. During the SAIL study many areas were considered and documented in order to define an MSK study window to ensure completeness of the patient referral data. This included ensuring that minimum GP registration periods and adequate follow-up times after diagnoses and procedures were applied. After addressing these challenges there may be a considerable reduction in size from the original dataset.

Event log data often requires pre-processing and sometimes advanced programming skill is needed due to data cleansing and data transformation requirements. Deciding on the case notion is an important step and in each study we demonstrate how to classify cases in order to create a particular view on the data. In the MIMIC-III study many data transformations were required in order to create an event log classified by directional ordered pairs of diseases.

It may be useful to split the event logs before process mining. In the MIMIC-III study separate event logs were generated using PROM data, as well as sex and age in order to identify any potential indicators of a health outcome.

Process mining in healthcare can be thought of as a chain, as seen in Figure 8.8. The chain begins in real life where people require healthcare services, the care is recorded in healthcare information systems that provide healthcare data, which can be better understood using process mining techniques.

Figure 8.8 The process mining in healthcare chain



The weakness of process mining is that it is dependent on the rest of the chain. In order to create a true learning health system, the outputs of process mining must be fed back into real life to complete the cycle. This concept is described by Coiera [409]. Due to the complexity of healthcare processes there is often the need for high involvement from clinical domain experts.

The studies have demonstrated different uses for process mining outputs. Current practice within the NHS is to quote episode statistics, such as those published for accident and emergency attendances and admissions [410]. As demonstrated in the MyPathway and SAIL studies, process mining techniques may also be used to generate statistics for specific services and providers. However these studies showed, when used to calculate time intervals between events (see Figure 7.23), process mining techniques were time consuming and required a high level of manual intervention. To automate these calculations would take considerable time and effort.

Where the real value lies is in the transparency it provides, made possible through a range of advanced visualisations. Process mining gives the ability to explore, in order to understand the entire end-to-end journey through the processes that lead to the statistics. Whilst statistics may be more effective at getting public attention, they can also be superficial and insufficient. When used alongside statistics, process mining techniques provide a deep and much more informed understanding of healthcare processes. More uses for process mining in healthcare include the generation of new hypotheses. Process mining tools and techniques provide healthcare researchers with the ability to quickly examine complex healthcare data from different perspectives. Conformance checking may be used to compare real healthcare data against clinical guidelines and best practice. It was also effectively used during the SAIL study to help identify data quality issues. Finally, process mining tools were used to draw a disease trajectory model from event log data. To create the event log, a series of data transformations was defined. Guntur Kusuma [329] extended this work to automate the programming effort, making disease trajectory model accessible to more people.

Process mining frameworks such as PM^2 provide detailed guidance on how to approach a process mining project and suggest different techniques to help approach some of the challenges mentioned above. Established tools such as ProM, Celonis and DISCO provide support and a number of advanced visualisations for a transparent, white-box approach.

When selecting healthcare datasets trade-offs have to be made between the speed of access, clinical relevance, data volume, data quality and level of support. These trade-offs will be more or less important, depending on the purpose of the dataset. Access to the MIMIC-III dataset was fast and because it was curated the data quality was high, though it was created in 2012, from the US and provided no access to the clinicians involved in the processes. Initial access to the MyPathway data was fast, it was from the UK and allowed for direct access to the clinicians and developers, though as it was uncurated, the data quality was extremely low. Finally, access to the SAIL data was slow, though it provided current, national-level population data, where insights were available from clinical experts working in a similar environment. As the purpose of the MIMIC-III and SAIL datasets were primarily to test the method then the fact that they were from a different country and in the case of MIMIC-III, from 2012 was not important. However, for the MyPathway study it was important to have current data from the UK and access to the clinical experts in order to identify potential indicators of a health outcome.

8.5 Conclusions

In this section, the overall conclusions are stated, a summary is presented on how each of the primary research questions have been answered, contributions to knowledge are listed, considerations for healthcare system development and process mining research are given and directions for possible future work are offered.

8.5.1 Overall conclusions

The work presented in this thesis provides evidence that routine healthcare data can be analysed using process mining techniques to provide insights that may benefit patients suffering with musculoskeletal conditions. However, there are many challenges (Section 3.5.1) that must be considered.

When selecting appropriate datasets, the data quality and level of pre-processing should be closely assessed to allow for the calculation of realistic analysis timescales and resources. In addition, the evaluation method must be considered at an early stage to allow for the results to be designed accordingly if statistical testing is required. Although the MyPathway results were found to be interesting, after clinical evaluation it was decided that the differences for the three possible indicators of a health outcome identified in the study were not strong enough and may be down to chance.

The complexities present in healthcare data are often better analysed using a range of different graphics rather than text-based database queries. Using the features available within process mining tools, event log data can be viewed at various levels of abstraction, via a multitude of different graphical representations, and data can be dynamically filtered and viewed from a variety of perspectives with a minimum amount of effort.

The outputs generated using process mining techniques can be used for a range of purposes. However, they are most useful for the generation of new hypotheses or to provide insights into healthcare data through a range of established visualisations. These visualisation may reveal important associations hidden within complex data.

Outputs may also be used to generate statistics. Though when calculating multiple time intervals through variable, complex pathways, extensive manual calculations are often required.

Alongside the many challenges of healthcare research there are also many opportunities. Clinical knowledge is rapidly changing and is often expressed in clinical guidelines, which can be used as the basis for conformance checking and process mining. Technical advances, such as mobile health applications provide a valuable source of data input and patient-centred care will generate new sources of data. Such data includes PROMs, which we have demonstrated can be linked to the EHR data to refine process mining. In summary, we can conclude that process mining in healthcare is going to become more widespread, more important and yield better results and the work presented in this thesis has helped to lay the foundations.

8.5.2 Answering the research questions

The three primary research questions presented in Section 1.3 have been broken down and answered using the three studies. The aim of this research was to explore the application of process mining to routine healthcare data in order to provide insights that may benefit patients suffering with musculoskeletal conditions. This section summarises how each of the research questions have been answered.

RQ1: Can historical data be analysed using process mining techniques in order to identify information that can be used to provide insights that may benefit patients suffering with musculoskeletal conditions? **Yes, provided that there is sufficient good quality data**[4]

Two out of the three studies in this research programme were successful in applying process mining techniques to historical data, in order to identify information that can

[4] See Section 3.5.1.2 for healthcare data quality issues.

be used to provide insights that may benefit patients suffering with musculoskeletal conditions. The MIMIC-III study demonstrated how process mining techniques can help to advance the study of disease trajectories using data stored in EHRs. Disease trajectories can be used to help identify early predictors for associated diseases. When our model was compared to Jensen's CVD trajectory model, gout was not found to be central condition and only visible when the model was adjusted to include the top 12 percent of diseases. After applying process mining techniques to the MyPathway patient data, the study was unable to identify possible indicators of a patient health outcome due to challenges in the dataset. However, potential changes to help improve data collection for future process mining studies were identified and discussed. Finally, the SAIL study showed how process mining techniques can be used to help create a knee pain surgery reference model using linked primary and secondary care data. The reference model was used to identify data cleansing issues by applying conformance checking techniques to the SAIL data. Process discovery techniques were then used to generate some useful TKR surgery statistics and answer important clinical questions.

RQ1.1 How can process mining techniques be applied to help discover unseen pathways and identify where healthcare processes deviate from what is expected by the healthcare professionals?

Process mining can be used to discover unseen pathways and identify where healthcare processes deviate from what is expected by healthcare professionals. The disease trajectory models created during the MIMIC-III study were discussed with a clinical domain expert who found the results to be interesting. In the majority of cases a logical explanation could be given for the associations between diseases. All three types of process mining were used in the SAIL study. Extensive data cleansing activities were applied to the data using conformance checking and enhancement techniques where deviations were easily identified and explored. After cleansing the data, process discovery techniques were used to generate some useful TKR surgery statistics and answer important clinical questions.

RQ1.2 Can process mining techniques provide healthcare professionals with information and estimations on healthcare utilisation that may benefit patients suffering with musculoskeletal conditions? **Yes, though other techniques may be more appropriate for calculating episode statistics.**

The methods and the results from the MIMIC-III and SAIL studies were evaluated by clinical domain experts and the results compared against those from similar studies. With the guidance from a process mining practitioner, the clinical experts found that process mining techniques allowed for fast and flexible time-ordered visualisations of the data. These visualisations were found to be particularly beneficial to exploratory analysis and to help understand end-to-end processes. Though for the generation of statistics, where multiple time measurements are needed for highly variable cases, alternative techniques may be more appropriate. The information created using these methods was considered to be useful and informative. Ad-hoc questions could be answered at low cost, especially when compared to traditional queries which are often error prone due to their level of complexity.

8.5.3 Contributions to knowledge and impact

The motivation behind this research was to contribute towards the process mining in healthcare community and also towards wider health informatics research. The focus was primarily on MSK conditions. Contributions have been made from the three studies to advance knowledge in these fields. The methodological insights gained from the MIMIC-III and SAIL studies may be of interest to process and data scientists working within the healthcare domain. The work carried out and the insights gained from the MyPathway study has contributed towards the design of a process-aware healthcare information system, therefore has advanced knowledge in the wider area of health informatics. These contributions are discussed in more detail below.

The work undertaken during the MIMIC-III study has pioneered the use of process mining techniques for disease trajectory modelling. The impact includes the publication of four peer-reviewed works. The first, entitled ‘A Process Mining Approach to Discovering Cardiovascular Disease Trajectories’, was presented as a poster at Medical Informatics Europe in 2018 [360]. Here, the work described in Chapter 5 was summarised and was the first study to 1) apply process mining techniques to the MIMIC-III data and 2) create disease trajectory models using process mining techniques. The method developed and described in Chapter 5 was then shared with and extended by a member of the process mining research group. This work lead to the co-authored publication entitled 'Process Mining of Disease Trajectories: A Feasibility Study' [361]. Following the feasibility study, the article entitled 'Process Mining of Disease Trajectories in MIMIC-III: A Case Study' [329] was published. Finally, a literature review on process mining of disease trajectories was co-authored and has been accepted for publication in 2021[362].

Innovation, brought about by work carried out in the MyPathway study has had impact on both patients attending physiotherapy at STHT and the future direction of a leading healthcare application development company. Diagrams created have been used to inform stakeholders and an outline for a new product road map has been designed. The work carried out in Chapter 6 has also contributed towards a new software development approach and the re-design of an existing healthcare information system to make it process-aware. This contribution is in direct support of a need recognised by Munoz-Gama et al. in their recent publication on the characteristics and challenges for process mining in healthcare [290].

Few studies in process mining exist where issues specific to MSK conditions are discussed or reported. A recent study by Remy et al. that demonstrated a method for event log generation by using a cohort of patients treated for low back pain [315] had some interesting ideas, though was lacking in detail. The author appropriately recognised the limitations of the study and stated the results must be interpreted with caution. The work presented in chapters 6 and 7 expands on the work of Remy by introducing new techniques to address important steps missing in [315] and to broaden the application by considering different challenges related to alternative source system architectures. The work presented in Chapter 6, proposes conformance checking as a technique that can be used to rigorously cleanse the data prior to process discovery and analysis. In Chapter 7, MSK specific considerations, such as how to identify patients for inclusion in the event log based on their condition are discussed. The selection process for cases into the event log, based on issues specific to healthcare information systems as opposed to a general management information systems, are discussed. An example of such issues may include whether and how to exclude patients that may be travelling or that have recently moved between locations that use a different healthcare information system.

Other insights gained from Chapter 7 centred around the importance of laterality. This is relevant to a range of medical conditions where people have a left and a right sided body part, for example knees. First is the importance of accurately recording and selecting the correct side and second is the challenges associated with modelling bilateral conditions.

A lack of studies using data from more than one hospital was reported by Erdogan and Tarhan [281]. Results from the SAIL study included statistics for knee pain surgery using linked data from the entire population of Wales. In addition, unlike the results reported by van Wanrooij [308] and Canjels et al. [314], the statistics produced from the work in Chapter 7 are generalisable. Also to date, this thesis contains the only published research that has applied process mining techniques to data from the SAIL databank.

8.5.4 Considerations for healthcare system development and process mining research

After reflecting upon the work within this thesis and evaluating different healthcare datasets and process mining tools, the following considerations should be taken into account by healthcare system developers and process mining researchers.

1. System developers should consider the automatic generation and maintenance of an event log when initially designing new healthcare information systems (Section 6.3.5.3). This would be beneficial in the following ways: 1) to reduce the amount of time needed during the data extraction stage of process mining projects; 2) to reduce the number of data transformations required during a process mining project; 3) to improve the event log data quality; 4) to allow healthcare system

developers more flexibility when selecting the type of database; and 5) to make process mining research possible, where it may previously not have been.

2. When designing architectures for healthcare information systems, system developers should consider the implications on process mining projects when data is stored in non-relational databases and warehouses.

3. Likewise, process mining researchers should allow for realistic, often extended timeframes when needing to extract event log data from non-relational databases. It may be necessary to undertake a feasibility study to help understand the data at a more detailed level before committing to large-scale projects.

4. Data integrity should be a major consideration. Healthcare researchers should be aware of the limitations of the data they use and ensure it is suitable to help answer particular research questions. For example, using hospital EHR information from within primary care databases is often unreliable as many data quality issues can exist. Therefore, it is preferable, where possible, to use data from linked primary and secondary care databases.

5. Process mining tools provide automated features to handle many of the more common event log transformations, for example, merging or splitting data items or dynamically changing the level of data abstraction. However, when complex transformations of the event logs are required, either due to data quality issues or to complex pre-processing rules, an advanced level of programming skill may be required.

6. Healthcare researchers should be aware of potential representational bias within the data. Bias may be caused by the purpose behind the recording of the data (e.g. financial reasons) or it may be caused by the reasons behind clinical or personal decisions (e.g. the reason why a patient underwent surgery may be based on the outcome of a previous procedure, or it may be based on a surgeons preferences).

7. Healthcare researchers should be aware of potential modelling and data quality issues when working with data that includes sidedness, as it is often important to know to which side of the body data and processes apply.

8.6 Future work

There is potential to build on the research presented in this thesis. Future work could make improvements in the following directions.

For each of the studies inferential statistics could be generated. This would involve a statistician undertaking a simulation study [411] during the planning stage, based on the descriptive statistics generated in this thesis, to ensure future analyses are adequately powered. The methods described in the three studies could then be applied to independent datasets in order to confirm and further explore the findings.

Further analysis could be carried out into the generation of MSK disease trajectory models using process mining techniques. Three areas to explore could include: 1) using the method described in Chapter 5 on linked primary and secondary care data may uncover a stronger or different association between gout and cardiovascular diseases, whilst also assessing the reproducibility of the method; 2) further refinements to the method in Chapter 5 could be made with regard to how and when repeating patient diagnoses are included in the event log; and 3) information on the average time interval between diseases may be considered. Apart from using process mining techniques, alternative methods could be used to explore the association between gout and CVD. This may require a prospective study or a retrospective analysis of a gout registry.

The statistics generated from the method described in Chapter 6 could be repeated with a higher volume of the Sheffield data, or with a different MyPathway dataset to confirm potential indicators of a health outcome. Analysing data for more than one speciality would add confidence to the results. The analysis could be extended by broadening the range of variables to include sensory data, such as step count or heart rate information, from different datasets.

Future directions for the SAIL study may include: 1) perform a sensitivity analysis by expanding the knee pain diagnosis code list from Chapter 7 to include secondary terms. 2) perform a large study using linked SAIL data to fully define a knee pain pathway. The knee pain surgery reference model from Chapter 7 could be extended to include all knee pain surgery types, as well as different knee pain events such as physiotherapy referrals, GP visits, tests and prescriptions.

For process mining to be used within healthcare services like the NHS, many of the methods and techniques described in this thesis would need to be automated. Ideally, when data is first input into healthcare systems an event log would automatically be updated. Versions of this event log could be used to generate dynamic custom dashboard views for use by healthcare professionals. These views may be tailored to include specific disease trajectory models or care pathways and be used to inform different local or national healthcare services.

Although the focus of this work is on MSK diseases, the method described in this thesis is generalisable and can be mapped to other clinical areas within the healthcare domain.

8.7 Final summary

Process mining is a relatively modern technique that, over the past 12 years has proven to be well equipped to manage many of the complexities associated with healthcare processes. It has been applied across many areas of healthcare, however, one understudied area is in MSK diseases. The work in this thesis has focused on how process mining techniques can be applied to routine data in order to provide insights that may benefit patients suffering with musculoskeletal conditions.

The MIMIC-III work presented in this thesis has pioneered the use of process mining techniques for disease trajectory modelling. Two disease trajectory models were produced, the first created by closely replicating Jensen’s data extraction and transformation rules and the second was a refinement to this model where repeating patient diagnoses were excluded. The increased level of sensitivity in the second model allowed for new relationships to be identified and when evaluated, the results were found to be more representative from a clinical perspective. Both models needed adjusting to include the top 12 percent of diseases before gout became visible.

However, to obtain a more accurate reflection of the associations both primary and secondary care data should be used.

Due to limitations with the MyPathway dataset, rather than the methodological approach, the three possible indicators of a poor health outcome identified during the study could not be confirmed by the clinical expert. However, many valuable insights were gained and considerations highlighted for future healthcare and process mining researchers along with developers of healthcare systems. This work has already had considerable impact on the efficiency of the application and subsequently the patients that use it, as well as on the future direction of the development company.

Process mining techniques were used to help create a knee pain surgery reference model before using the model to check conformance against the real-life linked primary and secondary care data from patients across the entire population of Wales. The validated data was used to create some knee pain surgery statistics and closely aligned to those from similar study that used non-process mining techniques. Findings were evaluated by an orthopaedic surgeon who found process mining to be an effective method for answering import clinical questions. This work has helped to fill a gap in the literature, where it was identified that little research had been done using data from more than one hospital. To date, no similar published studies exist where process mining techniques have been used to analyse knee pain surgery data.

The main lessons learnt from this programme of work have been summarised and translated into a list of considerations for reference by future healthcare system developers and process mining researchers in this field.

List of References

[1] S. Parsons, “The burden of musculoskeletal conditions,” *Medicine (Baltimore).*, vol. 42, no. 4, pp. 190–192, 2014.

[2] World Health Organization, “WHO global strategy on people- centred and integrated health services,” 2015.

[3] C. J. L. Murray *et al.*, “UK health performance: findings of the Global Burden of Disease Study 2010,” *Lancet*, vol. 381, no. 9871, pp. 997–1020, 2013.

[4] World Health Organisation, “Musculoskeletal conditions,” 2019. [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/musculoskeletal-conditions. [Accessed: 07-Jul-2020].

[5] L. March *et al.*, “Burden of disability due to musculoskeletal (MSK) disorders,” *Best Pract. Res. Clin. Rheumatol.*, vol. 28, no. 3, pp. 353–366, 2014.

[6] NHS England, “Musculoskeletal: Helping people with painful bone and joint conditions see the right person.” [Online]. Available: https://www.england.nhs.uk/ourwork/ltc-op-eolc/ltc-eolc/si-areas/musculoskeletal/#. [Accessed: 18-Jun-2019].

[7] Royal College of Physicians, “Underfunded. Underdoctored. Overstretched. The NHS in 2016.,” 2016.

[8] P. Yadav, M. Steinbach, V. Kumar, and G. Simon, “Mining Electronic Health Records (EHRs): A Survey,” *ACM Comput. Surv.*, vol. 50, no. 6, 2018.

[9] A. K. Jha *et al.*, “Use of Electronic Health Records in U.S. Hospitals,” *N. Engl. J. Med.*, vol. 360, pp. 1628–1638, 2009.

[10] R. Mans, W. M. P. van der Aalst, N. C. Russell, P. J. M. Bakker, and A. J. Moleman, “Process-aware information system development for the healthcare domain-consistency, reliability, and effectiveness,” in *International Conference on Business Process Management BPM 2009: Business Process Management Workshops*, 2009, pp. 635–646.

[11] W. van der Aalst *et al.*, “Process Mining Manifesto,” in *Business Process*

Management Workshops, 2011, pp. 169–194.

[12] W. van der Aalst, *Process Mining: Data Science in Action*, 2nd ed. Springer, 2016.

[13] D. R. Ferreira, *A Primer on Process Mining: Practical Skills with Python and Graphviz*. Switzerland: Springer, 2017.

[14] P. Gooch and A. Roudsari, "Computerization of workflows, guidelines, and care pathways: a review of implementation challenges for process-oriented health information systems," *JAMIA*, vol. 18, no. 6, pp. 738–748, 2011.

[15] T. Allweyer, *BPMN 2.0: Introduction to the Standard for Business Process Modeling*, 2nd ed. Books on Demand, 2016.

[16] M. L. van Eck, X. Lu, S. J. J. Leemans, and W. M. P. van der Aalst, "PM2 : A Process Mining Project Methodology," in *International Conference on Advanced Information Systems Engineering CAiSE 2015: Advanced Information Systems Engineering*, 2015, pp. 297–313.

[17] A. E. W. Johnson *et al.*, "MIMIC-III, a freely accessible critical care database.," *Sci. data*, vol. 3, p. 160035, 2016.

[18] A. Johnson, T. Pollard, and R. Mark, "The MIMIC-III Clinical Database," *PhysioNet*, 2016. [Online]. Available: https://doi.org/10.13026/C2XW26. [Accessed: 27-Apr-2017].

[19] Advanced Digital Innovation, "MyPathway® Communication Software for Healthcare," 2019. [Online]. Available: https://mypathway.healthcare/. [Accessed: 18-Jun-2019].

[20] SAIL, "SAIL DATABANK," 2020. [Online]. Available: https://saildatabank.com/. [Accessed: 25-Jun-2020].

[21] B. I. D. M. Center, "Beth Isreal Deaconess Medical Center," 2020. [Online]. Available: https://www.bidmc.org/. [Accessed: 25-Jun-2020].

[22] World Population Review, "Boston, Massachusetts Population, 2020," 2020. [Online]. Available: https://worldpopulationreview.com/us-cities/boston-population/. [Accessed: 25-Jun-2020].

[23] A. P. Kurniati, E. Rojas, D. Hogg, G. Hall, and O. A. Johnson, "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 J.*, vol. 25, no. 4, pp. 1878–1893, 2019.

[24] A. P. Kurniati, G. Hall, D. Hogg, and O. Johnson, “Process mining in oncology using the MIMIC-III dataset.,” in *International Conference on Data and Information Science. Physics Conference Series 971 012008*, 2018.

[25] A. Alharbi, A. Bulpitt, and O. Johnson, “Improving Pattern Detection in Healthcare Process Mining Using an Interval-Based Event Selection Method.,” in *International Conference on Business Process Management*, 2017, pp. 88–105.

[26] F. Fox, V. R. Aggarwal, H. Whelton, and O. Johnson, “A Data Quality Framework for Process Mining of Electronic Health Record Data,” in *IEEE International Conference on Healthcare Informatics (ICHI)*, 2018, pp. 12–21.

[27] STHT, “Our Hospitals,” 2020. [Online]. Available: https://worldpopulationreview.com/world-cities/sheffield-population. [Accessed: 24-Jun-2020].

[28] STHT, “PhysioWorks,” 2020. [Online]. Available: https://www.sth.nhs.uk/services/a-z-of-community-services?id=1. [Accessed: 24-Jun-2020].

[29] World Population Review, “Sheffield Population 2020,” 2020. [Online]. Available: https://worldpopulationreview.com/world-cities/sheffield-population/. [Accessed: 24-Jun-2020].

[30] SAIL, “SAIL Databank: 10 years of spearheading data privacy and research utility,” 2017. [Online]. Available: https://saildatabank.com/wp-content/uploads/SAIL_10_year_anniversary_brochure.pdf. [Accessed: 25-Jun-2020].

[31] Welsh Government, “Stats Wales: Population Estimates by local authority and year,” 2020. [Online]. Available: https://statswales.gov.wales/Catalogue/Population-and-Migration/Population/Estimates/Local-Authority/populationestimates-by-localauthority-year. [Accessed: 24-Jun-2020].

[32] A. R. Goldberger *et al.*, “PhysioBank, PhysioToolkit, and PhysioNet Components of a New Research Resource for Complex Physiologic Signals,” *Circulation*, vol. 101, no. 23, pp. e215–e220, 2000.

[33] MIMIC, “Requesting Access.” [Online]. Available: https://mimic.physionet.org/gettingstarted/access/. [Accessed: 27-Apr-2017].

[34] Advanced Digital Innovation, “Advanced Digital Innovation: An innovative technology company and consultancy,” 2019. [Online]. Available:

https://www.adi-uk.com/. [Accessed: 10-May-2019].

[35] e-Learning for Healthcare, “About the Data Security Awareness programme,” 2018. [Online]. Available: https://www.e-lfh.org.uk/programmes/data-security-awareness/. [Accessed: 25-Jun-2020].

[36] E. Coiera, *Guide to Health Informatics*, 3rd ed. 2015.

[37] E. Rojas, J. Munoz-Gama, M. Sepúlveda, and D. Capurro, “Process mining in healthcare: A literature review,” *J. Biomed. Inform.*, vol. 61, pp. 224–236, Jun. 2016.

[38] A. B. Jensen *et al.*, “Temporal disease trajectories condensed from population-wide registry data covering 6.2 million patients,” *Nat. Commun.*, vol. 5, no. May, pp. 1–10, 2014.

[39] NHS UK, “Primary Care.” [Online]. Available: https://www.england.nhs.uk/primary-care/. [Accessed: 16-Jan-2021].

[40] NHS Providers, “The NHS provider sector,” 2021. [Online]. Available: https://nhsproviders.org/topics/delivery-and-performance/the-nhs-provider-sector. [Accessed: 16-Jan-2021].

[41] Multiple Sclerosis Trust, “Care in the NHS,” 2018. [Online]. Available: https://mstrust.org.uk/a-z/care-in-the-nhs. [Accessed: 10-Jan-2022].

[42] T. Tulchinsky and E. Varavikova, “National Health Systems,” in *The New Public Health*, 3rd ed., Academic Press, 2014, pp. 643–728.

[43] World Health Organization, “Health system governance,” 2020. [Online]. Available: https://www.who.int/health-topics/health-systems-governance#tab=tab_1. [Accessed: 15-Jul-2020].

[44] NHS Digital, “Quality and Outcomes Framework,” 2019. [Online]. Available: http://content.digital.nhs.uk/qof. [Accessed: 01-Feb-2017].

[45] C. McCrorie, J. Benn, O. A. Johnson, and A. Scantlebury, “Staff expectations for the implementation of an electronic health record system: a qualitative study using normalisation process theory,” *BMC Med. Inform. Decis. Mak.*, vol. 19, no. 1, 2019.

[46] L. Nguyen, E. Bellucci, and L. T. Nguyen, “Electronic health records implementation: an evaluation of information system impact and contingency factors,” *Int. J. Med. Inform.*, vol. 83, no. 11, pp. 779–796, 2014.

[47] L. M. Van Swol, M. Kolb, and O. Asan, "'We are on the same page:' the importance of doctors EHR screen sharing for promoting shared information and collaborative decision-making," *J. Commun. Healthc.*, 2020.

[48] T. Highfill, "Do hospitals with electronic health records have lower costs? A systematic review and metaanalysis," *Int. J. Healthc. Manag.*, vol. 13, no. 1, pp. 65–71, 2020.

[49] C. S. Kruse, C. Kristof, B. Jones, E. Mitchell, and A. Martinez, "Barriers to Electronic Health Record Adoption: a Systematic Literature Review," *J. Med. Syst.*, vol. 40, 2016.

[50] M. A. Tutty, L. E. Carlasare, S. Lloyd, and C. A. Sinsky, "The complex case of EHRs: examining the factors impacting the EHR user experience," *J. Am. Med. Informatics Assoc.*, vol. 26, no. 7, pp. 673–677, 2019.

[51] EMIS Health, "EMIS Health," *EMIS Health Live: The New Normal*, 2020. [Online]. Available: https://www.emishealth.com/. [Accessed: 19-Aug-2020].

[52] The Phoenix Partnership (TPP), "What is SystmOne?," 2016. [Online]. Available: https://www.tpp-uk.com/products/systmone. [Accessed: 03-Feb-2017].

[53] R. J. Johnson, "A Comprehensive Review of an Electronic Health Record System Soon to Assume Market Ascendancy: EPIC," *J. Healthc. Commun.*, vol. 1, no. 4, p. 36, 2016.

[54] O. A. Johnson, H. S. F. Fraser, J. C. Wyatt, and J. D. Walley, "Electronic health records in the UK and USA," *Lancet*, vol. 384, no. 9947, p. 954, 2014.

[55] Department of Health, "Delivering 21st century IT support for the NHS: national strategic programme," London, 2002.

[56] NHS England, "Safer hospitals, safer wards: achieving an integrated digital care record.," 2013. [Online]. Available: http://www.england.nhs.uk/wp-content/uploads/2013/07/safer-hosp-safer-wards.pdf. [Accessed: 01-Feb-2018].

[57] NHS England, "NHS England offers Trusts over £100 million funding pot to set up centres of global digital excellence," 2016. [Online]. Available: https://www.england.nhs.uk/2016/08/centres-digital-excellence/. [Accessed: 16-Jul-2020].

[58] W. A. Khan, M. Hussain, K. Latif, M. Afzal, F. Ahmad, and L. Sungyoung, "Process interoperability in healthcare systems with dynamic semantic web

services," *Computing*, vol. 95, pp. 837–862, 2013.

[59] M. Prodel, V. Augusto, B. Jouaneton, L. Lamarsalle, and X. Xie, "Optimal Process Mining for Large and Complex Event Logs," *IEEE Trans. Autom. Sci. Eng.*, vol. 15, no. 3, pp. 1309–1325, 2018.

[60] R. Lenz and M. Reichert, "IT support for healthcare processes – premises, challenges, perspectives," *Data Knowl. Eng.*, vol. 61, no. 1, pp. 39–58, 2007.

[61] National Institute for Health and Care Excellence, "Improving health and social care through evidence-based guidance," 2020. [Online]. Available: https://www.nice.org.uk/. [Accessed: 09-Jul-2020].

[62] World Health Organization, "International Classification of Diseases (ICD) Information Sheet," 2020. [Online]. Available: https://www.who.int/classifications/icd/factsheet/en/. [Accessed: 15-Jul-2020].

[63] NHS Connecting for Health, *OPCS Classification of Interventions and Procedures Version 4.5. Volume I - Tabular index*. London: The Stationery Office, 2009.

[64] T. Benson, "The history of the Read codes: the inaugural James Read Memorial Lecture 2011," *Inform. Prim. Care*, vol. 19, pp. 173–82, 2011.

[65] NHS England, "SNOMED CT." [Online]. Available: https://www.england.nhs.uk/digitaltechnology/digital-primary-care/snomed-ct/. [Accessed: 15-Jul-2020].

[66] E. K. Pavalko, "Beyond Trajectories: Multiple Concepts for Analyzing Long-Term Process," in *Studying Aging and Social Change: Conceptual and Methodological Issues*, M. A. Hardy, Ed. Sage Publications, Inc., 1997, pp. 129–147.

[67] B. A. Pescosolido, "Patient Trajectories," *Wiley Blackwell Encycl. Heal. Illness, Behav. Soc.*, pp. 1770–1777, 2013.

[68] K. Sumida and C. P. Kovesdy, "Disease Trajectories Before ESRD: Implications for Clinical Management," *Semin. Nephrol.*, vol. 37, no. 2, pp. 132–143, 2017.

[69] S. A. Murray, M. Kendall, and K. Boyd, "Illness trajectories and palliative care," *BMJ*, vol. 330, no. 7498, pp. 1007–1011, 2005.

[70] A. Barabási and Z. N. Oltvai, "Network biology: understanding the cell's

functional organization," *Nat. Rev. Genet.*, vol. 5, pp. 101–113, 2004.

[71] I. Feldman, A. Rzhetsky, and D. Vitkup, "Network properties of genes harboring inherited disease mutations," *PNAS*, vol. 105, no. 11, pp. 4323–4328, 2008.

[72] A. Zanzoni, M. Soler-López, and P. Aloy, "A network medicine approach to human disease," *FEBS Lett.*, vol. 583, no. 11, pp. 1759–65, 2009.

[73] S. Jensen, P.B., Jense,. L.J. and Brunak, "Mining electronic health records: towards better research applications and clinical care," *Nat. Rev. Genet.*, vol. 13, no. 6, pp. 395–405, 2012.

[74] K. Shameer *et al.*, "A genome- and phenome-wide association study to identify genetic variants influencing platelet count and volume and their pleiotropic effects," *Hum. Genet.*, vol. 133, no. 1, pp. 95–109, 2014.

[75] A. B. Jensen, "Email to Samantha Sykes." 2017.

[76] B. K. Beaulieu-Jones, P. Orzechowski, and J. H. Moore, "Mapping patient trajectories using longitudinal extraction and deep learning in the MIMIC-III critical care database.," in *Pacific Symposium on Biocomputing 2018*, 2018, pp. 123–132.

[77] A. Giannoula, A. Gutierrez-Sacristán, Á. Bravo, F. Sanz, and L. I. Furlong, "Identifying temporal patterns in patient disease trajectories using dynamic time warping: A population-based study," *Sci. Rep.*, vol. 8, no. 1, p. 4216, Dec. 2018.

[78] X. Ji, S. A. Chun, and J. Geller, "Predicting Comorbid Conditions and Trajectories Using Social Health Records," *IEEE Trans. Nanobioscience*, vol. 15, no. 4, pp. 371–379, Jun. 2016.

[79] X. Wang, D. Sontag, and F. Wang, "Unsupervised learning of disease progression models," in *KDD '14: Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*, 2014, pp. 85–94.

[80] B. S. Glicksberg *et al.*, "Comparative analyses of population-scale phenomic data in electronic medical records reveal race-specific disease networks.," *Bioinformatics*, vol. 32, no. 12, pp. i101–i110, 2016.

[81] D. A. Hanauer and N. Ramakrishnan, "Modeling temporal relationships in large scale clinical associations.," *JAMIA*, vol. 20, no. 2, pp. 332–341, 2013.

[82] C. A. Hidalgo, N. Blumm, A. L. Barabási, and N. A. Christakis, “A Dynamic Network Approach for the Study of Human Phenotypes.,” *PLOS Comput. Biol.*, vol. 5, no. 4, 2009.

[83] C. A. Hidalgo, N. Blumm, A.-L. Barabási, and N. A. Christakis, “A Dynamic Network Approach for the Study of Human Phenotypes,” *PLoS Comput. Biol.*, vol. 5, no. 4, p. e1000353, 2009.

[84] K. Steinhaeuser and N. V. Chawla, “A network-based approach to understanding and predicting diseases.,” in *Social Computing and Behavioral Modeling*, 2009, pp. 1–8.

[85] J. C. T. Fairbank and P. B. Pynsent, “The Oswestry Disability Index,” *Spine (Phila. Pa. 1976).*, vol. 25, no. 22, pp. 2940–2953, 2000.

[86] Physiopedia, “12-Item Short Form Survey (SF-12),” 2020. [Online]. Available: https://www.physio-pedia.com/12-Item_Short_Form_Survey_(SF-12). [Accessed: 16-Jul-2020].

[87] R. Fitzpatrick, R. Davey, M. Buxton, and D. Jones, “Evaluating patient-based outcome measures for use in clinical trials.,” *Health Technol. Assess. (Rockv).*, vol. 2, no. 14, 1998.

[88] T. W. Concannon, “Editorials Can patient centered outcomes research improve healthcare?,” *BMJ*, vol. 351, 2015.

[89] M. Tew, K. Dalziel, P. Clarke, A. Smith, P. F. Choong, and M. Dowsey, “Patient-reported outcome measures (PROMs): can they be used to guide patient-centered care and optimize outcomes in total knee replacement?,” *Qual. Life Res.*, vol. 29, no. 12, pp. 3273–3283, 2020.

[90] N. H. I. Hjollund, “Fifteen Years’ Use of Patient-Reported Outcome Measures at the Group and Patient Levels: Trend Analysis,” *J. Med. Internet Res.*, vol. 21, no. 9, p. e15856, 2019.

[91] M. M. Holmes, G. Lewith, D. Newell, J. Field, and F. L. Bishop, “The impact of patient-reported outcome measures in clinical practice for pain: a systematic review,” *Qual. Life Res.*, vol. 26, no. 2, pp. 245–257, Feb. 2017.

[92] A. Coyle and C. Carpenter, “Patient experiences of their clinical management by Extended Scope Physiotherapists following attendance at an Orthopaedic Clinical Assessment Service.,” *Int. J. Pers. Centred Med.*, vol. 1, no. 3, 2011.

[93] M. R. Dunbar and Z. Ghogawala, “Patient-Reported Outcomes,” in *Quality Spine Care*, J. Ratliff, T. Albert, J. Cheng, and K. J., Eds. Cham: Springer

International Publishing, 2019, pp. 69–73.

[94] N. J. Devlin and J. Appleby, "Getting the most out of proms Putting health outcomes at the heart of NHS decision-making," 2010.

[95] O. Fennelly, C. Blake, F. Desmeules, D. Stokes, and C. Cunningham, "Patient-reported outcome measures in advanced musculoskeletal physiotherapy practice: a systematic review," *Musculoskeletal Care*, vol. 16, no. 1, pp. 188–208, Mar. 2018.

[96] P. N. Ramkumar, J. D. Harris, and P. C. Noble, "Patient-reported outcome measures after total knee arthroplasty," *Bone Joint Res.*, vol. 4, no. 7, pp. 120–127, Jul. 2015.

[97] L. T. Buller, A. S. McLawhorn, Y. Lee, M. Cross, S. Haas, and S. Lyman, "The Short Form KOOS, JR Is Valid for Revision Knee Arthroplasty," *J. Arthroplasty*, pp. 1–7, 2020.

[98] D. A. Revicki, D. Cella, R. D. Hays, J. A. Sloan, W. R. Lenderking, and N. K. Aaronson, "Responsiveness and minimal important differences for patient reported outcomes," *Health Qual. Life Outcomes*, vol. 4, no. 1, p. 70, Dec. 2006.

[99] S. Coretti, M. Ruggeri, and P. McNamee, "The minimum clinically important difference for EQ-5D index: a critical review," *Expert Rev. Pharmacoecon. Outcomes Res.*, vol. 14, no. 2, pp. 221–233, 2014.

[100] N. P. Hurst, P. Kind, D. Ruta, M. Hunter, and A. Stubbings, "Measuring health-related quality of life in rheumatoid arthritis: validity, responsiveness and reliability of EuroQol (EQ-5D).," *Rheumatology*, vol. 36, no. 5, pp. 551–559, 1997.

[101] C. B. Agborsangaya, D. Lau, M. Lahtinen, T. Cooke, and J. A. Johnson, "Health-related quality of life and healthcare utilization in multimorbidity: results of a cross-sectional survey," *Qual. Life Res.*, vol. 22, pp. 791–799, 2013.

[102] N. Parsons, X. L. Griffin, J. Achten, and M. L. Costa, "Outcome assessment after hip fracture IS EQ-5D THE ANSWER?," *Bone Jt. Res*, vol. 3, pp. 69–75, 2014.

[103] H. S. J. Picavet and N. Hoeymans, "Health related quality of life in multiple musculoskeletal diseases: SF-36 and EQ-5D in the DMC3 study," *Ann Rheum Dis*, vol. 63, pp. 723–729, 2004.

[104] EuroQol Research Foundation, “EQ-5D User Guide: Basic information on how to use the EQ-5D-3L instrument, version 6.” pp. 1–34, 2018.

[105] EuroQol Group, “EuroQol--a new facility for the measurement of health-related quality of life.,” *Health Policy (New. York).*, vol. 16, no. 3, pp. 199–208, 1990.

[106] E. R. Foundation, “EQ-5D User Guide: Basic information on how to use the EQ-5D-5L instrument, version 2.1.” pp. 1–26, 2015.

[107] M. F. Janssen, E. Birnie, J. A. Haagsma, and G. J. Bonsel, “Comparing the standard EQ-5D three-level system with a five-level version.,” *Value Heal.*, vol. 11, no. 2, pp. 275–284, 2008.

[108] M. F. Janssen, E. Birnie, and G. J. Bonsel, “Quantification of the level descriptors for the standard EQ-5D three-level system and a five-level version according to two methods.,” *Qual. Life Res.*, vol. 17, no. 3, pp. 463–473, 2008.

[109] M. F. Janssen *et al.*, “Measurement properties of the EQ-5D-5L compared to the EQ-5D-3L across eight patient groups: a multi-country study,” *Qual. Life Res.*, vol. 22, no. 7, pp. 1717–1727, 2013.

[110] B. Zamora, D. Parkin, Y. Feng, A. Bateman, M. Herdman, and N. Devlin, “New Methods for Analysing the Distribution of EQ-5D Observations,” *Office of Health Economics*. 2018.

[111] R. Rabin, F. de Charro, and A. Szende, *EQ-5D Value Sets*, vol. 2. Dordrecht: Springer Netherlands, 2007.

[112] D. Parkin, N. Devlin, N. Rice, and N. Devlin, “Statistical analysis of EQ-5D profiles: does the use of value sets bias inference?,” *Med. Decis. Mak.*, vol. 30, no. 5, pp. 556–565, 2010.

[113] N. Gutacker, C. Bojke, S. Daidone, N. Devlin, and A. Street, “Hospital Variation in Patient-Reported Outcomes at the Level of EQ-5D Dimensions: Evidence from England,” *Med. Decis. Mak.*, vol. 33, no. 6, pp. 804–818, 2013.

[114] N. J. Devlin, D. Parkin, and J. Browne, “Patient-reported outcome measures in the NHS: new methods for analysing and reporting EQ-5D data.,” *Heal. Econ.*, vol. 19, pp. 886–905, 2010.

[115] Y. Feng, D. Parkin, and N. J. Devlin, “Assessing the performance of the EQ-VAS in the NHS PROMs programme,” *Qual. Life Res.*, vol. 23, pp. 977–989,

2014.

[116] D. K. Whynes, “Correspondence between EQ-5D health state classifications and EQ VAS scores.,” *Health Qual. Life Outcomes*, vol. 6, no. 94, 2008.

[117] R. L. Kolotkin, R. D. Crosby, and G. R. Williams, “Integrating Anchor-Based and Distribution-Based Methods to Determine Clinically Meaningful Change in Obesity-Specific Quality of Life,” *Qual. Life Res.*, vol. 11, no. 7, p. 670, 2002.

[118] P. W. Stratford, J. M. Binkley, D. L. Riddle, and G. H. Guyatt, “Sensitivity to Change of the Roland-Morris Back Pain Questionnaire: Part 1,” *Phys. Ther.*, vol. 78, no. 11, pp. 1186–1196, Nov. 1998.

[119] Å. Jonsson, L. Orwelius, U. Dahlstrom, and M. Kristenson, “Evaluation of the usefulness of EQ-5D as a patient-reported outcome measure using the Paretian classification of health change among patients with chronic heart failure,” *J. Patient-Reported Outcomes*, vol. 4, no. 50, 2020.

[120] M. Asmirajanti, A. Y. Syuhaimie Hamid, and T. S. Hariyati, “Clinical care pathway strenghens interprofessional collaboration and quality of health service: a literature review,” *Enfermería Clínica*, vol. 28, pp. 240–244, Feb. 2018.

[121] L. de Bleser, R. Depreitre, K. de Waele, K. Vanhaecht, J. Vlayen, and W. Sermeus, “Defining pathways,” *J. Nurs. Manag.*, vol. 14, no. 7, pp. 553–563, 2006.

[122] K. W. Altman, “Improving Health Outcomes and Value with Care Pathways: The Otolaryngologist’s Role,” *Otolaryngol. Neck Surg.*, vol. 151, no. 4, pp. 527–529, 2014.

[123] A. den Hertog, K. Gliesche, J. Timm, B. Mu¨hlbauer, and S. Zebrowski, “Pathway-controlled fast-track rehabilitation after total knee arthroplasty: a randomized prospective clinical study evaluating the recovery pattern, drug consumption, and length of stay,” *Arch Orthop Trauma Surg*, vol. 132, no. 8, pp. 1153–1163, 2012.

[124] National Institute for Health and Care Excellence, “Osteoarthritis overview,” 2020. [Online]. Available: https://pathways.nice.org.uk/pathways/osteoarthritis. [Accessed: 23-Feb-2021].

[125] National Institute for Health and Care Excellence, “Managing osteoarthritis,” 2020. [Online]. Available:

https://pathways.nice.org.uk/pathways/osteoarthritis#content=view-index&path=view%3A/pathways/osteoarthritis/managing-osteoarthritis.xml. [Accessed: 23-Feb-2021].

[126] E. Aspland, D. Gartner, and P. Harper, “Clinical pathway modelling: a literature review,” *Heal. Syst.*, pp. 1–23, 2019.

[127] K. Button, F. Morgan, A. L. Weightman, and S. Jones, “Musculoskeletal care pathways for adults with hip and knee pain referred for specialist opinion: a systematic review,” *BMJ Open*, vol. 9, no. 9, p. e027874, 2019.

[128] D. P. Gwynne-Jones, L. R. Hutton, K. M. Stout, and J. H. Abbott, “The joint clinic: managing excess demand for hip and knee osteoarthritis referrals using a new physiotherapy-led outpatient service.,” *J. Arthroplasty*, vol. 33, no. 4, pp. 983–987, 2018.

[129] N. Parfitt, A. Smeatham, J. Timperley, M. Hubble, and G. Gie, “Direct listing for total hip replacement (THR) by primary care physiotherapists.,” *Clin. Gov. An Int. J.*, vol. 17, no. 3, pp. 210–216, 2012.

[130] Huang, Z., W. Dong, H. Duan, and H. Li, “Similarity Measure Between Patient Traces for Clinical Pathway Analysis: Problem, Method, and Applications,” *IEEE J. Biomed. Heal. Informatics*, vol. 18, no. 1, pp. 4–14, 2014.

[131] F. Caron, J. Vanthienen, and B. Baesens, “Healthcare Analytics: Examining the Diagnosis–treatment Cycle,” *Procedia Technol.*, vol. 9, pp. 996–1004, 2013.

[132] Arthritis Research UK, “Musculoskeletal Health: A public health approach,” 2014.

[133] World Health Organization, “Global burden of disease.” [Online]. Available: http://www.who.int/topics/global_burden_of_disease/en/. [Accessed: 17-Jul-2017].

[134] C. Murray, T. Vos, R. Lozano, and et al, “Disability-adjusted Life Years (DALYs) for 291 Diseases and Injuries in 21 Regions, 1990-2010: A systematic analysis for the Global Burden of Disease Study 2010,” *Lancet*, vol. 380, pp. 2197–2223, 2012.

[135] T. Vos, A. Flaxman, M. Naghavi, and et al, “Years Lived with Disability (YLDs) for 1160 Sequelae of 289 Diseases and Injuries 1990-2010: A systematic analysis for the Global Burden of Disease Study 2010,” *Lancet*, vol. 380, pp. 2163–2196, 2012.

[136] R. Lozano *et al.*, "Global and regional mortality from 235 causes of death for 20 age groups in 1990 and 2010: a systematic analysis for the Global Burden of Disease Study 2010," *Lancet*, vol. 380, no. 9859, pp. 2095–2128, Dec. 2012.

[137] DoH, "Reference Costs 2015-16," 2016.

[138] Arthritis Research UK, "Osteoarthritis in general practice: Data and perspectives," 2013.

[139] National Institute for Health and Care Excellence, "Osteoarthritis: care and management," 2014. [Online]. Available: https://www.nice.org.uk/guidance/CG177/chapter/1-Recommendations. [Accessed: 02-Feb-2017].

[140] J. H. Klippel and P. Dieppe, "Disorders of the musculoskeletal system in Rheumatoid Arthritis. Clinical features of rheumatoid arthritis: early progressive and late disease.," *Pract. Rheumatol.*, pp. 169–182, 1997.

[141] J. E. Collins, N. J. Katz, E. E. Dervan, and E. Losina, "Trajectories and risk profiles of pain in persons with radiographic, symptomatic knee osteoarthritis: data from the osteoarthritis initiative.," *Osteoarthr. Cartil.*, vol. 22, no. 5, pp. 622–630, 2014.

[142] E. Nicholls, E. Thomas, T. E. van der Windt, P. R. Croft, and G. Peat, "Pain trajectorygroups in persons with, or at high risk of, knee osteoarthritis: findings from the Knee Clinical AssessmentStudy and the Osteoarthritis Initiative.," *Osteoarthr. Cartil.*, vol. 22, pp. 2041–50, 2014.

[143] J. F. M. Holla *et al.*, "Three trajectories of activity limitations in early symptomatic knee osteoarthritis: a 5-year follow-up study," *Ann. Rheum. Dis.*, vol. 73, no. 7, pp. 1369–1375, Jul. 2014.

[144] S. J. Bartlett, S. M. Ling, N. E. Mayo, S. C. Scott, and C. O. Bingham, "Identifying common trajectories of joint space narrowing over two years in knee osteoarthritis," *Arthritis Care Res. (Hoboken).*, vol. 63, no. 12, pp. 1722–1728, Dec. 2011.

[145] P. Nair, J. Ting, H. I. Keen, and P. G. Conaghan, "Arthritis in older adults," in *Brocklehursts Textbook of Geriatric Medicine and Gerontology*, 8th ed., H. M. Fillit, K. Rockwood, and J. Young, Eds. Elsevier, 2017, pp. 552–564.

[146] National Osteoporosis Society, "Osteoarthritis and Osteoporosis," 2016. [Online]. Available: https://nos.org.uk/about-osteoporosis/other-bone-conditions/osteoarthritis-and-

osteoporosis/?gclid=Cj0KCQjwwLHLBRDEARIsAN1A1Q55HA1awtmwR81-SgYyD9WZdigKWgcVX86c20S181zIxdOBRFYDxPUaAr1uEALw_wcB. [Accessed: 17-Jul-2017].

[147] D. T. Felson *et al.*, "Osteoarthritis: new insights. Part 1: the disease and its risk factors.," *Ann. Intern. Med.*, vol. 133, no. 8, pp. 635–646, 2000.

[148] M. D. Sewell, K. Rosendahl, and D. M. Eastwood, "Developmental dysplasia of the hip," *BMJ*, vol. 339, no. 4, pp. 1242–1248, 2009.

[149] R. Wittenauer, L. Smith, and K. Aden, "Priority Medicines for Europe and the World " A Public Health Approach to Innovation " Update on 2004 Background Paper Background Paper 6.12 Osteoarthritis," *World Heal. Organ.*, pp. 1–31, 2013.

[150] M. T. Hirschmann and W. Müller, "Complex function of the knee joint: the current understanding of the knee," *Knee Surgery, Sport. Traumatol. Arthrosc.*, vol. 23, no. 10, pp. 2780–2788, Oct. 2015.

[151] Z. Meyler, "Knee Anatomy," *ARTHRITIS-health*, 2018.

[152] D. J. Culliford *et al.*, "The lifetime risk of total hip and knee arthroplasty: results from the UK general practice research database," *Osteoarthr. Cartil.*, vol. 20, no. 6, pp. 519–524, 2012.

[153] National Institute for Health and Care Excellence, "Knee pain - assessment," 2017. [Online]. Available: https://cks.nice.org.uk/topics/knee-pain-assessment/. [Accessed: 25-Dec-2020].

[154] D. C. Crawford, L. E. Miller, and J. E. Block, "Conservative management of symptomatic knee osteoarthritis: a flawed strategy?," *Orthop. Rev. (Pavia).*, vol. 5, no. 1, p. 2, Feb. 2013.

[155] T. E. McAlindon *et al.*, "OARSI guidelines for the non-surgical management of knee osteoarthritis," *Osteoarthr. Cartil.*, vol. 22, no. 3, pp. 363–388, 2014.

[156] O. O. Babatunde, J. L. Jordan, D. A. Van der Windt, J. C. Hill, N. E. Foster, and J. Protheroe, "Effective treatment options for musculoskeletal pain in primary care: A systematic overview of current evidence," *PLoS One*, vol. 12, no. 6, p. e0178621, Jun. 2017.

[157] National Institute for Health and Care Excellence, "Osteoarthritis: When should I suspect a diagnosis of osteoarthritis?," 2018. [Online]. Available: https://cks.nice.org.uk/osteoarthritis#!diagnosisSub. [Accessed: 04-Feb-2020].

[158] C. Reyes, K. M. Leyland, G. Peat, C. Cooper, N. K. Arden, and D. Prieto-Alhambra, “Association Between Overweight and Obesity and Risk of Clinically Diagnosed Knee, Hip, and Hand Osteoarthritis: A Population-Based Cohort Study,” *Arthritis Rheumatol.*, vol. 68, no. 8, pp. 1869–1875, Aug. 2016.

[159] L. Jiang *et al.*, “Body mass index and susceptibility to knee osteoarthritis: A systematic review and meta-analysis,” *Jt. Bone Spine*, vol. 79, no. 3, pp. 291–297, May 2012.

[160] M. Reijman *et al.*, “Body mass index associated with onset and progression of osteoarthritis of the knee but not of the hip: The Rotterdam Study,” *Ann. Rheum. Dis.*, vol. 66, no. 2, pp. 158–162, Aug. 2006.

[161] M. Grotle, K. B. Hagen, B. Natvig, F. A. Dahl, and T. K. Kvien, “Obesity and osteoarthritis in knee, hip and/or hand: An epidemiological study in the general population with 10 years follow-up,” *BMC Musculoskelet. Disord.*, vol. 9, no. 1, p. 132, Dec. 2008.

[162] D. Prieto-Alhambra, A. Judge, M. K. Javaid, C. Cooper, A. Diez-Perez, and N. K. Arden, “Incidence and risk factors for clinically diagnosed knee, hip and hand osteoarthritis: influences of age, gender and osteoarthritis affecting other joints,” *Ann. Rheum. Dis.*, vol. 73, no. 9, pp. 1659–1664, Sep. 2014.

[163] L. Murphy *et al.*, “Lifetime risk of symptomatic knee osteoarthritis,” *Arthritis Rheum.*, vol. 59, no. 9, pp. 1207–1213, Sep. 2008.

[164] NHS, “Overview - Arthroscopy,” 2020. [Online]. Available: https://www.nhs.uk/conditions/arthroscopy/. [Accessed: 15-Dec-2020].

[165] K. F. Pajalic, A. Turkiewicz, and M. Englund, “Update on the risks of complications after knee arthroscopy,” *BMC Musculoskelet. Disord.*, vol. 19, 2018.

[166] M. A. Adelani, A. H. S. Harris, T. R. Bowe, and N. J. Giori, “Arthroscopy for Knee Osteoarthritis Has Not Decreased After a Clinical Trial,” *Clin. Orthop. Relat. Res.*, vol. 474, no. 2, pp. 489–494, Feb. 2016.

[167] D. H. Howard, “Trends in the Use of Knee Arthroscopy in Adults,” *JAMA Intern Med.*, vol. 178, no. 11, pp. 1557–1558, 2018.

[168] R. A. C. Siemieniuk *et al.*, “Arthroscopic surgery for degenerative knee arthritis and meniscal tears: a clinical practice guideline,” *BMJ*, vol. 357, p. j1982, 2017.

[169] J. B. Thorlund, C. B. Juhl, E. M. Roos, and L. S. Lohmander, “Arthroscopic surgery for degenerative knee: systematic review and meta-analysis of benefits and harms,” *BMJ*, vol. 350, 2015.

[170] A. R. Winter, J. E. Collins, and J. N. Katz, “The likelihood of total knee arthroplasty following arthroscopic surgery for osteoarthritis: a systematic review,” *BMC Musculoskelet. Disord.*, vol. 18, no. 1, p. 408, 2017.

[171] S. R. Bollen, “Is arthroscopy of the knee completely useless?,” *bone Jt. J.*, vol. 97–B, no. 12, pp. 1591–1592, 2015.

[172] D. J. Culliford, J. Maskell, D. J. Beard, D. W. Murray, A. J. Price, and N. K. Arden, “Temporal trends in hip and knee replacement in the United Kingdom,” *J. Bone Joint Surg. Br.*, vol. 92–B, no. 1, pp. 130–135, Jan. 2010.

[173] A. J. Price *et al.*, “Knee replacement,” *Lancet*, vol. 392, no. 10158, pp. 1672–82, 2018.

[174] National Joint Registry, “National Joint Registry 17th Annual Report 2020,” 2020.

[175] S. Morgan, “Total Knee Replacement.” [Online]. Available: https://www.samermorgan.com/total-knee-replacement/. [Accessed: 16-Dec-2020].

[176] NHS, “Overview - Knee replacement,” 2019. [Online]. Available: https://www.nhs.uk/conditions/knee-replacement/. [Accessed: 16-Dec-2020].

[177] K. Harris *et al.*, “Systematic review of measurement properties of patient-reported outcome measures used in patients undergoing hip and knee arthroplasty,” *Patient Relat Outcome Meas.*, vol. 7, pp. 101–108, 2016.

[178] J. T. Evans, R. W. Walker, J. P. Evans, A. W. Blom, A. Sayers, and M. R. Whitehouse, “How long does a knee replacement last? A systematic review and meta-analysis of case series and national registry reports with more than 15 years of follow-up,” *Lancet*, vol. 393, pp. 655–63, 2019.

[179] A. Postler, C. Lützner, F. Beyer, E. Tille, and J. Lützner, “Analysis of Total Knee Arthroplasty revision causes,” *BMC Musculoskelet. Disord.*, vol. 19, no. 55, pp. 1–6, 2018.

[180] B. Zmistowski, J. A. Karam, J. B. Durinka, D. S. Casper, and J. Parvizi, “Periprosthetic joint infection increases the risk of one-year mortality,” *J Bone Jt. Surg Am.*, vol. 95, no. 24, pp. 2177–84, 2013.

[181] R. Walker-Santiago, J. D. Tegethoff, W. M. Ralston, and J. A. Keeney, “Revision Total Knee Arthroplasty in Young Patients: Higher Early Reoperation and Rerevision,” *J. Arthroplasty*, vol. 36, no. 2, pp. 653–656, 2021.

[182] E. Zachwieja, J. Perez, W. M. Hardaker, B. Levine, and N. Sheth, “Manipulation Under Anesthesia and Stiffness After Total Knee Arthroplasty,” *J. Bone Jt. Surgery, Inc.*, vol. 6, no. 4, p. e2, 2018.

[183] Arthritis Research UK, “State of Musculoskeletal Health 2017: Arthritis & other musculoskeletal conditions in numbers.,” 2017. [Online]. Available: https://www.versusarthritis.org/. [Accessed: 25-Aug-2017].

[184] Chartered Institute for Professional Development, “Absence management. Annual survey report 2014.,” 2014.

[185] M. Urwin *et al.*, “Estimating the burden of musculoskeletal disorders in the community: the comparative prevalence of symptoms at different anatomical sites, and the relation to social deprivation,” *Ann. Rheum. Dis.*, vol. 57, no. 11, pp. 649–655, 1998.

[186] I. A. Bernstein, Q. Malik, S. Carville, and S. Ward, “Low back pain and sciatica: summary of NICE guidance,” *BMJ*, vol. 356, p. i6748, 2017.

[187] National Institute for Health and Care Excellence, “NICE publishes updated advice on treating low back pain,” 2016. [Online]. Available: https://www.nice.org.uk/news/article/nice-publishes-updated-advice-on-treating-low-back-pain. [Accessed: 10-Jul-2020].

[188] D. Hoy *et al.*, “The global burden of low back pain: estimates from the Global Burden of Disease 2010 study,” *Ann Rheum Dis*, vol. 73, pp. 1470–1476, 2014.

[189] I. Heuch, I. Heuch, K. Hagen, and J.-A. Zwart, “Body Mass Index as a Risk Factor for Developing Chronic Low Back Pain,” *Spine (Phila. Pa. 1976).*, vol. 38, no. 2, pp. 133–139, Jan. 2013.

[190] M. B. Pinheiro *et al.*, “Symptoms of Depression and Risk of New Episodes of Low Back Pain: A Systematic Review and Meta-Analysis,” *Arthritis Care Res. (Hoboken).*, vol. 67, no. 11, pp. 1591–1603, Nov. 2015.

[191] Alkherayf, “Daily smoking and lower back pain in adult Canadians: the Canadian Community Health Survey,” *J. Pain Res.*, vol. 3, p. 155, Aug. 2010.

[192] Arthritis Research UK, “Musculoskeletal conditions and multimorbidity,”

2017.

[193] C. Maher, M. Underwood, and R. Buchbinder, “Non-specific low back pain,” *Lancet*, vol. 389, no. 10070, pp. 736–747, 2017.

[194] C. Greenough, “The National Back Pain Pathway.,” *NHS England*, 2016. [Online]. Available: https://www.england.nhs.uk/blog/charles-greenough/#. [Accessed: 25-Aug-2017].

[195] T. Pascart and F. Lioté, “Gout: state of the art after a decade of developments,” *Rheumatology*, vol. 58, no. 1, pp. 27–44, 2019.

[196] M. Dehlin, L. Jacobsson, and E. Roddy, “Global epidemiology of gout: prevalence, incidence, treatment patterns and risk factors,” *Nat. Rev. Rheumatol.*, vol. 16, pp. 380–390, 2020.

[197] R. Soskind, D. T. Abazia, and M. Barna Bridgeman, “Updates on the treatment of gout, including a review of updated treatment guidelines and use of small molecule therapies for difficult-to-treat gout and gout flares,” *Expert Opin. Pharmacother.*, vol. 18, no. 11, pp. 1115–1125, 2017.

[198] M. Chen-Xu, C. Yokose, S. K. Rai, M. H. Pillinger, and H. K. Choi, “Contemporary Prevalence of Gout and Hyperuricemia in the United States and Decadal Trends: The National Health and Nutrition Examination Survey 2007-2016,” *Arthritis Rheumatol.*, vol. 71, no. 6, pp. 991–999, 2019.

[199] T. Neogi *et al.*, “2015 Gout Classification Criteria: An American College of Rheumatology/European League Against Rheumatism Collaborative Initiative,” *Arthritis Rheumatol.*, vol. 67, no. 10, pp. 2557–2568, 2015.

[200] National Institute for Health and Care Excellence, “Gout,” 2020. [Online]. Available: https://cks.nice.org.uk/gout#!scenario. [Accessed: 10-Jul-2020].

[201] E. F. Codd, “A Relational Model of Data for Large Shared Data Banks,” *Commun. ACM*, vol. 13, no. 6, pp. 377–387, 1970.

[202] R. Reinsch, “Distributed database for SAA,” *IBM Syst. J.*, vol. 27, no. 3, pp. 362–389, 1988.

[203] IBM, “NoSQL Databases,” 2019. [Online]. Available: https://www.ibm.com/cloud/learn/nosql-databases. [Accessed: 03-Sep-2020].

[204] M. Fowler and J. Highsmith, “The agile manifesto,” in *Software development,* 2001, pp. 28–32.

[205] R. Gall, “Different types of NoSQL databases and when to use them,” *Packt*, 2019. [Online]. Available: https://hub.packtpub.com/different-types-of-nosql-databases-and-when-to-use-them/. [Accessed: 03-Sep-2020].

[206] D. Chauhan and K. L. Bansal, “Using the Advantages of NOSQL: A Case Study on MongoDB,” *Int. J. Recent Innov. Trends Comput. Commun.*, vol. 5, no. 2, pp. 90–93, 2017.

[207] C. O’Neil and R. Schutt, *Doing Data Science*. O’Reilly Media, Inc., 2014.

[208] I. D. Constantiou and J. Kallinikos, “New Games, New Rules: Big Data and the Changing Context of Strategy,” *J. Inf. Technol.*, vol. 30, no. 1, pp. 44–57, Mar. 2015.

[209] V. Huser and J. J. Cimino, “Impending Challenges for the Use of Big Data,” *Int. J. Radiat. Oncol.*, vol. 95, no. 3, pp. 890–894, 2016.

[210] T. Huang, L. Lan, X. Fang, P. An, J. Min, and F. Wang, “Promises and Challenges of Big Data Computing in Health Sciences,” *Big Data Res.*, vol. 2, no. 1, pp. 2–11, 2015.

[211] T. B. Murdoch and A. S. Detsky, “The inevitable application of big data to health care,” *JAMA*, vol. 309, no. 13, pp. 1351–1352, 2013.

[212] J. vom Brocke *et al.*, “Process Science: The Interdisciplinary Study of Continuous Change,” 2021.

[213] D. C. Wynn and P. J. Clarkson, “Process models in design and development,” *Res. Eng. Des.*, vol. 29, pp. 161–202, 2017.

[214] S. M. R. Beheshti *et al.*, *Process Analytics*. Springer International Publishing, 2016.

[215] T. Crocker, O. A. Johnson, and S. F. King, “TOWARDS A FORMALISATION OF CARE PATHWAYS TO EMBODY GOOD PRACTICE IN HEALTHCARE,” in *Proceedings of eGovernment Workshop (eGOV07)*, 2007.

[216] National Institute for Health and Care Excellence, “How NICE clinical guidelines are developed: an overview for stakeholders, the public and the NHS,” *The guidelines manual*, 2012. [Online]. Available: https://www.nice.org.uk/process/pmg6/resources/how-nice-clinical-guidelines-are-developed-an-overview-for-stakeholders-the-public-and-the-nhs-2549708893/chapter/about-nice-guidance. [Accessed: 02-Sep-2020].

[217] R. Anupindi, S. Deshmukh, J. Van Mieghem, and E. Zemel, *Managing business process flows*, 2nd ed. New Jersey: Prentice Hall, 1999.

[218] G. Blokdyk, *Data Flow Diagram A Complete Guide - 2020 Edition*. 5STARCooks, 2020.

[219] G. Booch, J. Rumbaugh, and I. Jacobson, *The Unified Modeling Language User Guide*, 2nd ed. Addison-Wesley Professional, 2005.

[220] T. Allweyer and D. Allweyer, *BPMN 2.0: Introduction to the Standard for Business Process Modeling*, 2nd ed. Books on Demand GmbH, Norderstedt, 2010.

[221] The Object Management Group, “Object Management Group,” 2020. [Online]. Available: https://www.omg.org/. [Accessed: 03-Sep-2020].

[222] S. Bennett, S. Mcrobb, and R. Farmer, *Object-Oriented Systems Analysis and Design Using UML*, 3rd ed. McGraw-Hill Education, 2006.

[223] N. Russell, A. Ter, W. Van der Aalst, and P. Wohed, “On the Suitability of UML 2.0 Activity Diagrams for Business Process Modelling,” in *Third Asia-Pacific Conference on Conceptual Modelling (APCCM 2005)*, 2005.

[224] C. V. Geambasu, “BPMN VS. UML ACTIVITY DIAGRAM FOR BUSINESS PROCESS MODELING,” *J. Account. Manag. Inf. Syst.*, vol. 11, no. 4, pp. 637–651, 2012.

[225] B. Silver, *BPMN Method & Style*, 2nd ed. Cody-Cassidy Press, 2011.

[226] IEEE, “Carl Adam Petri,” *2008 Computer Pioneer Award Recipient*, 2008. [Online]. Available: https://mycomputer.computer.org/web/awards/pioneer-carl-petri. [Accessed: 03-Sep-2020].

[227] J. Desel, W. Reisig, and G. Rozenberg, Eds., *Lectures on Concurrency and Petri Nets*, vol. 3098. Berlin, Heidelberg: Springer Berlin Heidelberg, 2004.

[228] Y. Huang, Y. Weng, and M. Zhou, “Design of Traffic Safety Control Systems for Emergency Vehicle Preemption Using Timed Petri Nets,” *IEEE Trans. Intell. Transp. Syst.*, vol. 16, no. 4, pp. 2113–2120, 2015.

[229] P. J. Haas, *Stochastic Petri Nets: Modelling, Stability, Simulation*. Springer, 2006.

[230] Pnml.org, “Welcome on PNML.org: Tool construction,” 2015. [Online].

Available: http://www.pnml.org/tools.php. [Accessed: 05-Jan-2021].

[231] Agerwala, “Special Feature: Putting Petri Nets to Work,” *Computer (Long. Beach. Calif).*, vol. 12, no. 12, pp. 85–94, Dec. 1979.

[232] W. J. Thong and M. A. Ameedeen, “A Survey of Petri Net Tools,” in *Advanced Computer and Communication Engineering Technology. Lecture Notes in Electrical Engineering, vol 315*, 2015, pp. 537–551.

[233] U. of H. TGI group, “Welcome to the Petri Nets World,” 2020. [Online]. Available: http://www.informatik.uni-hamburg.de/TGI/PetriNets/index.php. [Accessed: 05-Jan-2021].

[234] WoPeD, “Workflow Petri Net Designer.” [Online]. Available: https://woped.dhbw-karlsruhe.de/. [Accessed: 04-Jan-2021].

[235] W. van der Aalst, “Process Mining: Bridging Not Only Data and Processes, but Also Industry and Academia.” [Online]. Available: https://www.celonis.com/blog/process-mining-bridging-not-only-data-and-processes-but-also-industry-and-academia/. [Accessed: 04-Dec-2019].

[236] W. van der Aalst, T. Weijters, and L. Maruster, “Workflow mining: discovering process models from event logs,” *IEEE Trans. Knowl. Data Eng.*, vol. 16, no. 9, pp. 1128–1142, 2004.

[237] J. E. Cook, “Process discovery and validation through event-data analysis.,” University of Colorado, 1996.

[238] R. Agrawal, D. Gunopulos, and F. Leymann, “Mining process models from workflow logs.,” in *International Conference on Extending Database Technology*, 1998, pp. 467–483.

[239] W. M. P. van der Aalst, B. F. van Dongen, J. Herbst, L. Maruster, G. Schimm, and A. J. M. M. Weijters, “Workflow mining: A survey of issues and approaches,” *Data Knowl. Eng.*, vol. 47, no. 2, pp. 237–267, Nov. 2003.

[240] M. Ghasemi and D. Amyot, “From event logs to goals: a systematic literature review of goal-oriented process mining,” *Requir. Eng.*, vol. 25, no. 1, pp. 67–93, 2020.

[241] A. Corallo, M. Lazoi, and F. Striani, “Process mining and industrial applications: A systematic literature review,” *Knowl. Process Manag.*, 2020.

[242] R. Ahmed, M. Faizan, and A. I. Burney, “Process Mining in Data Science: A Literature Review,” in *MACS 2019 - 13th International Conference on*

Mathematics, Actuarial Science, Computer Science and Statistics, 2019, p. R. Ahmed, M. Faizan, A. I. Burney.

[243] N. M. El-Gharib and D. Amyot, “Process mining for cloud-based applications: A systematic literature review,” in *Proceedings - 2019 IEEE 27th International Requirements Engineering Conference Workshops, REW 2019*, 2019, pp. 34–43.

[244] N. F. Farid, M. de Kamps, and O. A. Johnson, “Process Mining in Frail Elderly Care: A Literature Review,” in *Proceedings of the 12th International Joint Conference on Biomedical Engineering Systems and Technologies - Volume 5: HEALTHINF.*, 2019, pp. 332–339.

[245] W. van der Aalst *et al.*, “Process Mining Manifesto,” in *Lecture Notes in Business Information Processing*, vol. 99 LNBIP, no. PART 1, 2012, pp. 169–194.

[246] C. Alvarez *et al.*, “Discovering role interaction models in the Emergency Room using Process Mining,” *J. Biomed. Inform.*, vol. 78, pp. 60–77, 2018.

[247] Z. Zhou, Y. Wang, and L. Li, “Process mining based modeling and analysis of workflows in clinical care - A case study in a chicago outpatient clinic,” in *Proceedings of the 11th IEEE International Conference on Networking, Sensing and Control*, 2014, pp. 590–595.

[248] A. J. M. M. Weijters, W. M. P. van der Aalst, and A. K. A. de Medeiros, “Process mining with the heuristics miner-algorithm,” 2006.

[249] C. W. Gunther and W. M. P. van der Aalst, “Fuzzy mining: Adaptive process simplification based on multi-perspective metrics,” in *BPM 2007*, 2007, pp. 328–343.

[250] S. J. J. Leemans, D. Fahland, and W. van der Aalst, “Discovering Block-Structured Process Models from Event Logs - A Constructive Approach,” in *International Conference on Applications and Theory of Petri Nets and Concurrency PETRI NETS 2013: Application and Theory of Petri Nets and Concurrency*, 2013, pp. 311–329.

[251] J. C. A. M. Buijs, B. F. van Dongen, and W. M. P. van der Aalst, *On the Role of Fitness, Precision, Generalization and Simplicity in Process Discovery.*, vol. 7565. Springer Berlin Heidelberg, 2012.

[252] F. Mannhardt, M. de Leoni, and H. A. Reijers, “The Multi-perspective Process Explorer,” in *BPM (Demos), CEUR Workshop Proceedings*, 2015, pp. 130–134.

[253] G. Greco, A. Guzzo, L. Pontieri, and D. Saccà, "Discovering expressive process models by clustering log traces," *IEEE Trans. Knowl. Data Eng.*, vol. 18, no. 8, pp. 1010–1027, 2006.

[254] K. Kirchner, N. Herzberg, A. Rogge-Solti, and M. Weske, "Embedding Conformance Checking in a Process Intelligence System in Hospital Environments.," in *BPM' 2012 Proceedings of the 2012 international conference on Process Support and Knowledge Representation in Health Care*, 2012, pp. 126–139.

[255] J. Y. Zhou, "Process mining: acquiring objective process information for healthcare process management with the CRISP-DM framework," Eindhoven University of Technology, 2009.

[256] Eindhoven University of Technology, "Process Mining in Healthcare." Future Learn, 2017.

[257] E. Batista and A. Solanas, "Process Mining in Healthcare: A Systematic Review," in *2018 9th International Conference on Information, Intelligence, Systems and Applications (IISA)*, 2018, pp. 1–6.

[258] L. Wang, Y. Du, and L. Qi, "Efficient Deviation Detection Between a Process Model and Event Logs," *IEEE-CAA J. Autom. Sin.*, vol. 6, no. 6, pp. 1352–1364, 2019.

[259] A. Adriansyah, B. Dongen, and W. van der Aalst, "Towards Robust Conformance Checking," in *Lecture Notes in Business Information Processing*, 2010, pp. 122–133.

[260] D. Han and Y. Tian, "Analysis and Application of Transition Systems Based on Petri Nets and Relation Matrices to Business Process Management," *Math. Probl. Eng.*, pp. 1–18, 2020.

[261] M. Estañol, J. Munoz-Gama, J. Carmona, and E. Teniente, "Conformance checking in UML artifact-centric business process models," *Softw. Syst. Model.*, vol. 18, pp. 2531–2555, 2019.

[262] D. Dakic, S. Sladojevic, T. Lolic, and D. Stefanovic, "Process Mining Possibilities and Challenges: A Case Study," in *SISY 2019 - IEEE 17th International Symposium on Intelligent Systems and Informatics*, 2019, pp. 161–166.

[263] P. Badakhshan and A. Alibabaei, *Using Process Mining for Process Analysis Improvement in Pre-hospital Emergency*, vol. 35. Springer, Cham, 2020.

[264] P. Chapman *et al.*, *CRISP-DM 1.0: Step-by-step data mining guide*. SPSS, 2000, 2000.

[265] A. Azevedo and M. F. Santos, “KDD, SEMMA and CRISP-DM: a parallel overview.,” in *IADIS European Conf. Data Mining. Vol. 8*, 2008, pp. 182–185.

[266] M. Bozkaya, J. Gabriels, and J. M. Van der Werf, “Process Diagnostics: A Method Based on Process Mining,” in *International Conference on Information, Process, and Knowledge Management, 2009. eKNOW ’09.*, 2009.

[267] M. L. van Eck, X. Lu, S. J. J. Leemans, and W. M. P. van der Aalst, “PM 2 : a Process Mining Project Methodology.,” in *International Conference on Advanced Information Systems Engineering (CAiSE) 2015. Lecture Notes in Computer Science, vol 9097*, 2015, pp. 297–313.

[268] A. Syamsiyah *et al.*, “Business Information Systems: 20th International Conference, BIS 2017.,” in *Business Process Comparison: A Methodology and Case Study*, 2017, pp. 253–267.

[269] S. Suriadi, M. T. Wynn, C. Ouyang, A. H. M. Ter Hofstede, and N. J. Van Dijk, “Understanding Process Behaviours in a Large Insurance Company in Australia: A Case Study,” in *CAiSE 2013: Advanced Information Systems Engineering*, 2013, pp. 449–464.

[270] B. F. van Dongen, A. K. A. de Medeiros, H. M. W. Verbeek, A. J. M. M. Weijters, and W. M. P. van der Aalst, “The ProM Framework: A New Era in Process Mining Tool Support,” in *Application and Theory of Petri Nets 2005*, no. i, 2005, pp. 444–454.

[271] C. W. Günther and A. Rozinat, “Disco: Discover Your Processes,” in *BPM 2012*, 2012, pp. 40–44.

[272] J. Kiefer and M. Precht, “USAGE OF PROCESS MINING IN THE ‘OFFER TO PRODUCTION PROCESS’ OF A CONTRACT MANUFACTURER FOR CAST COMPONENTS,” in *Multi Conference on Computer Science and Information Systems, MCCSIS 2019 - Proceedings of the International Conferences on Big Data Analytics, Data Mining and Computational Intelligence 2019 and Theory and Practice in Modern Computing 2019*, 2019, pp. 103–110.

[273] Eindhoven University of Technology, “ProM Tools.” [Online]. Available: http://www.promtools.org/doku.php. [Accessed: 20-Jul-2020].

[274] C. W. Günther and W. M. P. van der Aalst, “Fuzzy Mining – Adaptive Process Simplification Based on Multi-perspective Metrics,” in *Business Process Management. BPM 2007. Lecture Notes in Computer Science, vol 4714*, 2007.

[275] T. The Process Mining Group, “MXML (Mining eXtensible Markup Language),” *Process Mining: research, tools, application*, 2016. [Online]. Available: http://www.processmining.org/logs/mxml. [Accessed: 21-May-2019].

[276] IEEE, *Standard for eXtensible Event Stream (XES) for Achieving Interoperability in Event Logs and Event Streams*, vol. 1849. 2016.

[277] W. Yang, Q. Su, and S. Qiang, “Process Mining for Clinical Pathway Literature Review and Future Directions,” in *11th Int. Conf. Service Syst. & Service Management*, 2014, pp. 1–5.

[278] E. Rojas, M. Arias, and M. Sepúlveda, “Clinical Processes and Its Data, What Can We Do with Them?,” in *Proceedings of the International Conference on Health Informatics*, 2015, pp. 642–647.

[279] M. Ghasemi and D. Amyot, “Process mining in healthcare: a systematised literature review,” *Int. J. Electron. Healthc.*, vol. 9, no. 1, p. 60, 2016.

[280] T. Erdoğan and A. Tarhan, “Process Mining for Healthcare Process Analytics,” in *2016 Joint Conference of the International Workshop on Software Measurement and the International Conference on Software Process and Product Measurement (IWSM-MENSURA)*, 2016, pp. 125–130.

[281] T. G. Erdogan and A. Tarhan, “Systematic Mapping of Process Mining Studies in Healthcare,” *IEEE Access*, vol. 6, pp. 24543–24567, 2018.

[282] M. R. Dallagassa, C. dos Santos Garcia, E. E. Scalabrin, S. O. Ioshii, and D. R. Carvalho, “Opportunities and challenges for applying process mining in healthcare: a systematic mapping study,” *J. Ambient Intell. Humaniz. Comput.*, 2021.

[283] A. P. Kurniati, O. Johnson, D. Hogg, and G. Hall, “Process Mining in Oncology : a Literature Review,” in *6th International Conference on Information Communication and Management (ICICM)*, 2016, pp. 291–297.

[284] G. P. Kusuma, M. Hall, C. P. Gale, and O. A. Johnson, “Process Mining in Cardiology: A Literature Review,” *Int. J. Biosci. Biochem. Bioinforma.*, vol. 8, no. 4, pp. 226–236, 2018.

[285] R. Williams, E. Rojas, N. Peek, and O. A. Johnson, “Process Mining in Primary Care: A Literature Review,” *Stud. Health Technol. Inform.*, vol. 247, pp. 376–380, 2018.

[286] E. Helm, A. M. Lin, D. Baumgartner, A. C. Lin, and J. Küng, “Towards the Use of Standardized Terms in Clinical Case Studies for Process Mining in Healthcare,” *Int. J. Environ. Res. Public Health*, vol. 17, no. 4, p. 1348, 2020.

[287] R. S. Mans, W. M. P. van der Aalst, and R. J. . Vanwersch, *Process Mining in Healthcare: Evaluating and Exploiting Operation Healthcare Processes*. Springer Cham, 2015.

[288] FutureLearn, “Process Mining in Healthcare.” FutureLearn, 2020.

[289] “Process-Oriented Data Science for Healthcare.” 2020.

[290] J. Munoz-Gama *et al.*, “Process Mining for Healthcare: Characteristics and Challenges,” *J. Biomed. Inform.*, vol. 127, p. 103994, 2022.

[291] P. Homayounfar, “Process mining challenges in hospital information systems,” in *Computer Science and Information Systems (FedCSIS), 2012 Federated Conference on.*, 2012, pp. 1135–1140.

[292] M. Song, C. W. Günther, and W. Van der Aalst, “Trace Clustering in Process Mining,” in *Lecture Notes in Business Information Processing*, 2008, pp. 109–120.

[293] B. F. A. Hompes, J. C. A. M. Buijs, P. M. van der Aalst, W. M. P., Dixit, and J. Buurman, “Discovering Deviating Cases and Process Variants Using Trace Clustering,” in *27th Benelux Conference on Artificial Intelligence*, 2015.

[294] L. Vanbrabant, N. Martin, K. Ramaekers, and K. Braekers, “Quality of input data in emergency department simulations: Framework and assessment techniques,” *Simul. Model. Pract. Theory*, vol. 91, pp. 83–101, 2019.

[295] N. G. Weiskopf and C. Weng, “Methods and dimensions of electronic health record data quality assessment: enabling reuse for clinical research,” *JAMIA*, vol. 20, no. 1, pp. 144–151, 2013.

[296] S. J. van Zelst, F. Mannhardt, M. de Leoni, and A. Koschmider, “Event abstraction in process mining: literature review and taxonomy,” *Granul. Comput.*, vol. 6, pp. 719–736, 2021.

[297] NHS Health Research Authority, “Guidance for using patient data,” 2020. [Online]. Available: https://www.hra.nhs.uk/covid-19-research/guidance-

using-patient-data/. [Accessed: 11-Sep-2020].

[298] HM Government, "Human Rights Act 1998," 2022. [Online]. Available: https://www.legislation.gov.uk/ukpga/1998/42/contents. [Accessed: 21-Feb-2022].

[299] HM Government, "National Health Service Act 2006," 2022. [Online]. Available: https://www.legislation.gov.uk/ukpga/2006/41/contents. [Accessed: 21-Feb-2022].

[300] HM Government, "Health and Social Care Act 2012." [Online]. Available: https://www.legislation.gov.uk/ukpga/2012/7/contents/enacted. [Accessed: 21-Feb-2022].

[301] Information Commissioner's Office, "Guide to the UK General Data Protection Regulation (UK GDPR)." [Online]. Available: https://ico.org.uk/for-organisations/guide-to-data-protection/guide-to-the-general-data-protection-regulation-gdpr/. [Accessed: 21-Feb-2022].

[302] W. W. Lowrance, "Access to collections of data and materials for health research : a report to the Medical Research Council and the Wellcome Trust," 2006.

[303] I. Beerepoot, X. Lu, I. van de Weerd, and H. A. Reijers, "Seeing the signs of workarounds: a mixed-methods approach to the detection of nurses' process deviations," in *54th Hawaii International Conference on System Sciences (HICSS)*, 2021, pp. 1–10.

[304] T. The Process Mining Group, "Papers About Process Mining in Healthcare," 2014. [Online]. Available: http://www.processmining.org/health/papers. [Accessed: 03-Feb-2021].

[305] SAIL Databank, "Projects Using SAIL," 2021. [Online]. Available: https://saildatabank.com/wp-content/https://saildatabank.com/saildata/projects-using-sail/. [Accessed: 03-Feb-2021].

[306] T. Aagaard, H. Lund, and C. Juhl, "Optimizing literature search in systematic reviews – are MEDLINE, EMBASE and CENTRAL enough for identifying effect studies within the area of musculoskeletal disorders?," *BMC Med Res Methodol.*, vol. 16, no. 161, 2016.

[307] Phd Assistance, "Is Google A Right Source For Identifying My Literature Review?," 2020. [Online]. Available: https://www.phdassistance.com/blog/is-google-a-right-source-for-identifying-my-literature-review/. [Accessed: 10-

Feb-2021].

[308] K. P. A. van Wanrooij, “Patient Careflow Discovery,” Utrecht University, 2012.

[309] Z. Valero-Ramon, G. Ibanez-Sanchez, V. Traver, L. Marc0-Ruiz, and C. Fernandez-Llatas, “Towards Perceptual Spaces for Empowering Ergonomy in Workplaces by using Interactive Process Mining,” in *Transforming Ergonomics with Personalized Health and Intelligent Workplaces*, IOS Press, 2019, pp. 85–100.

[310] P. Asgharia, A. Masoud, and H. H. S. Javadi, “Internet of Things applications: A systematic review,” *Comput. Networks*, vol. 148, pp. 241–261, 2019.

[311] V. A. Petrushin, “Hidden markov models: Fundamentals and applications,” in *Online Symposium for Electronics Engineer*, 2000.

[312] W. S. Noble, “What is a support vector machine?,” *Nat. Biotechnol.*, vol. 24, pp. 1565–1567, 2006.

[313] Z. Zhang, “Artificial Neural Network,” in *Multivariate Time Series Analysis in Climate and Environmental Research*, Springer, Cham, 2018, pp. 1–35.

[314] K. F. Canjels, M. S. V. Imkamp, T. A. E. J. Boymans, and R. J. B. Vanwersch, “Unraveling and improving the interorganizational arthrosis care process at Maastricht UMC+: an illustration of an innovative, combined application of data and process mining.,” in *BPM (Industry Forum) 2019*, 2019, pp. 178–189.

[315] W. M. Remy S., Pufahl L., Sachs J.P., Böttinger E., “Event Log Generation in a Health System: A Case Study,” in *Business Process Management. BPM 2020. Lecture Notes in Computer Science, vol 12168*, 2020, pp. 505–522.

[316] L. Corr and J. Stagnitto, *Agile data warehouse design: Collaborative dimensional modeling, from whiteboard to star schema*. DecisionOne Press, 2011.

[317] J. J. Koorn *et al.*, “Bringing Rigor to the Qualitative Evaluation of Process Mining Findings: An Analysis and a Proposal,” in *3rd International Conference on Process Mining (ICPM)*, 2021, pp. 120–127.

[318] A. Suchenia, P. Wisniewski, and A. Ligeza, “Overview of Verification Tools for Business Process Models,” in *2017 Federated Conference on Computer Science and Information Systems*, 2017, pp. 295–302.

[319] G. Kang, L. Yang, and L. Zhang, “Verification of behavioral soundness for artifact-centric business process model with synchronizations,” *FutureGenerationComputerSystems*, vol. 98, pp. 503–511, 2019.

[320] H. Verbeek, T. Basten, and W. van der Aalst, “Diagnosing Workflow Processes using Woflan,” *Comput. J.*, vol. 44, pp. 246–279, 2001.

[321] R. Laue and A. Awad, “Visualization of Business Process Modeling Anti Patterns,” in *Electronic Communications of the EASST , 25*, 2010.

[322] N. Trcka, W. M. van der Aalst, and N. Sidorova, “Data-flow anti-patterns: Discovering data-flow errors in workflows,” in *Advanced Information Systems Engineering*, 2009, pp. 425–439.

[323] A. Koenig, “Patterns and antipatterns,” *J. Object-Oriented Program.*, vol. 8, no. 1, pp. 46–48, 1995.

[324] T. Jager, “Dynamic Modeling for Uptake and Effects of Chemicals,” in *Marine Ecotoxicology Current Knowledge and Future Issues*, J. Blasco, P. M. Chapman, O. Campana, and M. Hampel, Eds. Academic Press, 2016, pp. 71–98.

[325] A. K. Alves de Medeiros *et al.*, “An Outlook on Semantic Business Process Mining and Monitoring,” in *On the Move to Meaningful Internet Systems 2007: OTM 2007 Workshops. OTM 2007. Lecture Notes in Computer Science, vol 4806*, 2007, pp. 1244–1255.

[326] F. Mannhardt and D. Blinde, “Analyzing the Trajectories of Patients with Sepsis using Process Mining,” in *RADAR+ EMISA@ CAiSE*, 2017, pp. 72–80.

[327] D. Berrar, “Cross-Validation,” in *Encyclopedia of Bioinformatics and Computational Biology, Volume 1*, Elsevier, 2019, pp. 542–545.

[328] J. Motl and P. Kordík, “Stratified Cross-Validation on Multiple Columns,” in *2021 IEEE 33rd International Conference on Tools with Artificial Intelligence (ICTAI)*, 2021, pp. 26–31.

[329] G. Kusuma, A. Kurniati, C. D. McInerney, M. Hall, C. P. Gale, and O. Johnson, “Process Mining of Disease Trajectories in MIMIC-III: A Case Study,” in *Lecture Notes in Business Information Processing. 2nd International Conference on Process Mining (ICPM 2020)*, 2020.

[330] I. H. Witten, E. Frank, M. A. Hall, and C. Pal, *Data Mining : Practical Machine Learning Tools and Techniques*, 4th ed. Elsevier Science &

Technology, 2016.

[331] S. E. Bleeker *et al.*, “External validation is necessary in prediction research: A clinical example,” *J. Clin. Epidemiol.*, vol. 56, no. 9, pp. 826–832, 2003.

[332] J. Futoma, M. Simons, T. Panch, F. Doshi-Velez, and L. A. Celi, “The myth of generalisability in clinical research and machine learning in health care,” *Lancet Digit. Heal.*, vol. 2, no. 9, pp. e489–e492, 2020.

[333] S. Y. Ho, K. Phua, L. Wong, and W. W. B. Goh, “Extensions of the External Validation for Checking Learned Model Interpretability and Generalizability,” *Patterns*, vol. 1, no. 8, p. 100129, 2020.

[334] G. Upton and I. Cook, *Understanding Statistics*. Oxford University Press, 2003.

[335] R. B. Davis and K. J. Mukamal, “Hypothesis Testing: means,” *Circulation*, vol. 114, no. 10, pp. 1078–1082, 2006.

[336] A. N. Christopher, “Drawing Conclusions From Data: Descriptive Statistics, Inferential Statistics, and Hypothesis Testing,” in *Interpreting and Using Statistics in Psychological Research*, Sage Publications, Inc., 2017, pp. 145–183.

[337] I. P. Cardenas, M. Espinoza, and H. Armas-Aguirre, J. Aguirre-Mayorga, “Security of the information model on process mining: case study of the surgery block,” in *2021 Congreso Internacional de Innovación y Tendencias en Ingeniería (CONIITI)*, 2021, pp. 1–5.

[338] R. Lira *et al.*, “Process-Oriented Feedback through Process Mining for Surgical Procedures in Medical Training: The Ultrasound-Guided Central Venous Catheter Placement Case,” *Int. J. Environ. Res. Public Heal.*, vol. 16, no. 11, p. 1877, 2019.

[339] T. Gurgen Erdogan and A. Tarhan, “A Goal-Driven Evaluation Method Based On Process Mining for Healthcare Processes,” *Appl. Sci.*, vol. 8, no. 6, p. 894, 2018.

[340] E. Benevento, P. M. Dixit, M. F. Sani, D. Aloini, and W. M. van der Aalst, “Evaluating the Effectiveness of Interactive Process Discovery in Healthcare: A Case Study,” in *International conference on business process management*, 2019, pp. 508–519.

[341] L. Canensi, G. Leonardi, S. Montani, and P. Terenziani, “Multi-level interactive medical process mining,” in *Conference on Artificial Intelligence*

in Medicine in Europe, 2017, pp. 256–260.

[342] F. G. Fox, “Applying Process-Oriented Data Science to Dentistry,” University of Leeds, 2019.

[343] M. h. Loxton, “Process Mining in Healthcare Quality Improvement,” 2014. [Online]. Available: https://www.researchgate.net/profile/Matthew-Loxton/publication/303048589_Process_Mining_Case_Studies_in_Healthcare/links/5736199b08ae298602e09ffc/Process-Mining-Case-Studies-in-Healthcare.pdf. [Accessed: 07-Feb-2022].

[344] R. Clay-Williams, J. Hounsgaard, and E. Hollnagel, “Where the rubber meets the road: using FRAM to align work-as-imagined with work-as-done when implementing clinical guidelines,” *Implement. Sci.*, vol. 10, no. 1, pp. 1–8, 2015.

[345] C. K. Christian *et al.*, “A prospective study of patient safety in the operating room,” *Surgery*, vol. 139, no. 2, pp. 159–73, 2006.

[346] K. Catchpole, D. M. Neyens, J. Abernathy, D. Allison, A. Joseph, and S. T. Reeves, “Framework for direct observation of performance and safety in healthcare,” *BMJ Qual Saf*, vol. 26, no. 12, pp. 1015–1021, 2017.

[347] M. Song and W. Van der Aalst, “Supporting process mining by showing events at a glance.,” in *Seventeenth Annual Workshop on Information Technologies and Systems (WITS’07)*, 2007, pp. 139–145.

[348] J. Rumbaugh, I. Jacobson, and G. Booch, *Unified Modeling Language Reference Manual*, Second. Pearson Higher Education, 2004.

[349] RStudio Team, “RStudio: Integrated Development for R.,” 2015. [Online]. Available: http://www.rstudio.com/. [Accessed: 06-Dec-2019].

[350] Microsoft Corporation, “Microsoft Excel.” 2016.

[351] R. Ghawi, “Process Discovery using Inductive Miner and Decomposition,” 2016.

[352] J. Sundbøll *et al.*, “Positive predictive value of cardiovascular diagnoses in the Danish National Patient Registry: a validation study.,” *BMJ Open*, vol. 6, no. 11, p. e012832, 2017.

[353] World Health Organization, “Classifications,” 2016. [Online]. Available: http://www.who.int/classifications/icd/en/. [Accessed: 27-Apr-2017].

[354] The PostgreSQL Global Development Group, “PostgreSQL 9.6.3 Documentation,” 2017. [Online]. Available: https://www.postgresql.org/docs/9.6/static/release-9-6.html. [Accessed: 16-May-2017].

[355] mongoDB, “What is MongoDB?,” 2019. [Online]. Available: https://www.mongodb.com/what-is-mongodb?jmp=search&utm_source=google&utm_campaign=GS_EMEA_United Kingdom_Search_Brand_Atlas_Desktop&utm_keyword=mongo database&utm_device=c&utm_network=g&utm_medium=cpc&utm_creative=335278754570&utm_matchtype=e&_bt=33527875. [Accessed: 08-May-2019].

[356] D. Crockfor, “Introducing JSON.” [Online]. Available: https://www.json.org/. [Accessed: 08-May-2019].

[357] S. Bramwell, “ADI presentation.” 2018.

[358] NHS Wales Informatics Service, “Admitted Patient Care Data Set (APC Ds),” *NHS WALES DATA DICTIONARY*, 2019. [Online]. Available: http://www.datadictionary.wales.nhs.uk/index.html#!WordDocuments/admittedpatientcaredatasetapcds.htm. [Accessed: 28-Sep-2020].

[359] S. Gkisser, *Predictive Inference: An Introduction*. New York: Chapman and Hall/CRC, 1993.

[360] S. J. Sykes, S. R. Kingsbury, P. G. Conaghan, M. P. Rodriguez, P. D. Baxter, and O.A. Johnson, “A Process Mining Approach to Discovering Cardiovascular Disease Trajectories,” in *Medical Informatics Europe*, 2018.

[361] G. Kusuma, S. Sykes, C. McInerney, and O. Johnson, “Process Mining of Disease Trajectories: A Feasibility Study,” in *Proceedings of the 13th International Joint Conference on Biomedical Engineering Systems and Technologies*, 2020, pp. 705–712.

[362] G. P. KUSUMA *et al.*, “Process Mining of Disease Trajectories: A Literature Review,” in *Medical Informatics Europe 2021*, 2021.

[363] MIMIC, “MIMIC-III Critical Care Database.” [Online]. Available: https://mimic.physionet.org/about/mimic/. [Accessed: 13-Mar-2021].

[364] The PostgreSQL Global Development Group, “pgAdmin 4 v1.1 Released!,” 2017. [Online]. Available: https://www.postgresql.org/about/news/1711/. [Accessed: 16-May-2017].

[365] Tomp, “what is the exactly relationship between the diagnoses priority and the treatment in MIMIC-III?,” 2016. [Online]. Available: https://opendata.stackexchange.com/questions/9335/what-is-the-exactly-relationship-between-the-diagnoses-priority-and-the-treatmen or Appendix B. [Accessed: 12-May-2017].

[366] Python Software Foundation, “Python 3.5.3 documentation.” [Online]. Available: https://docs.python.org/3.5/. [Accessed: 16-May-2017].

[367] Python Software Foundation, “spyder 2.3.9.,” 2017. [Online]. Available: https://pypi.python.org/pypi/spyder/2.3.9. [Accessed: 16-May-2017].

[368] Oracle, “Java JDK 8u20 Update Release Notes,” 2017. [Online]. Available: http://www.oracle.com/technetwork/java/javase/8u20-relnotes-2257729.html. [Accessed: 16-May-2017].

[369] The Eclipse Foundation, “Eclipse Neon 3 Packages,” 2017. [Online]. Available: http://www.eclipse.org/downloads/packages/. [Accessed: 16-May-2017].

[370] M. B. Dixon, “QSEE Technologies.” [Online]. Available: https://www.leedsbeckett.ac.uk/qsee/. [Accessed: 16-May-2017].

[371] ICD.Codes, “ICD-9-CM Converter.” [Online]. Available: https://icd.codes/convert/icd9-to-icd10-cm. [Accessed: 16-May-2017].

[372] B. J. Marafino, J. M. Davies, N. S. Bardach, M. L. Dean, and R. A. Dudley, “N-gram support vector machines for scalable procedure and diagnosis classification, with applications to clinical free text data from the intensive care unit,” *JAMIA*, vol. 21, no. 5, pp. 871–875, 2014.

[373] A. Chib and S. H. Lin, “Theoretical Advancements in mHealth: A Systematic Review of Mobile Apps,” *J. Health Commun.*, vol. 23, no. 10–11, pp. 909–955, Nov. 2018.

[374] A. J. Litterini and C. M. Wilson, “Measuring and Quantifying Outcomes,” in *Physical Activity and Rehibilitation in Life-threatening Illness*, Oxon, UK: Taylor and Francis, 2021.

[375] BusinessCloud, “£2m funding to expand NHS use of MyPathway HealthTech app,” 2020. [Online]. Available: https://businesscloud.co.uk/remote-healthtech-app-adi-raises-2m-to-expand-nhs-use/. [Accessed: 22-Mar-2021].

[376] UK Research and Innovation, “AIM-FORE : Adaptive Inter-Domain Models for Optimisation and Real-Time Evaluation.” [Online]. Available:

https://gtr.ukri.org/projects?ref=133522. [Accessed: 22-Mar-2021].

[377] H. Wilson, "Interview with Clinical Service Manager and Enhanced Role Physiotherapist." Sheffield, 2019.

[378] G. Sowden, J. C. Hill, L. Morso, Q. Louw, and N. E. Foster, "Advancing practice for back pain through stratified care (STarT Back)," *Brazilian J. Phys. Ther.*, vol. 22, no. 4, pp. 255–264, 2018.

[379] J. Fairbank, "Revised Oswestry Disability questionnaire.," *Spine (Phila PA 1976)*, vol. 25, no. 19, p. 2552, 2000.

[380] G. Peat, E. Thomas, R. Duncan, L. Wood, E. Hay, and P. Croft, "Clinical classification criteria for knee osteoarthritis: performance in the general population and primary care," *Ann Rheum Dis.*, vol. 65, no. 10, pp. 1363–1367, 2006.

[381] R. Bender and S. Lange, "Adjusting for multiple testing—when and how?," *J. Clin. Epidemiol.*, vol. 54, no. 4, pp. 343–349, 2001.

[382] G. E. Bekkering *et al.*, "Prognostic Factors for Low Back Pain in Patients Referred for Physiotherapy: Comparing Outcomes and Varying Modeling Techniques," *Spine (Phila. Pa. 1976).*, vol. 30, no. 16, pp. 1881–1886, 2005.

[383] Karmarkar, T. D. *et al.*, "A Fresh Perspective on a Familiar Problem: Examining Disparities in Knee Osteoarthritis Using a Markov Model," *Med. Care*, vol. 55, no. 12, pp. 993–1000, 2017.

[384] R. Chester, L. Shepstone, H. Daniell, D. Sweeting, J. Lewis, and C. Jerosch-Herold, "Predicting response to physiotherapy treatment for musculoskeletal shoulder pain: a systematic review," *BMC Musculoskelet. Disord.*, vol. 14, no. 203, 2013.

[385] J. Kaur and K. S. Mann, "AI based HealthCare Platform for Real Time, Predictive and Prescriptive Analytics using Reactive Programming," in *10th International Conference on Computer and Electrical Engineering*, 2018, vol. 933.

[386] A. Benis, N. Harel, R. B. Barkan, E. Srulovici, and C. Key, "Patterns of Patients' Interactions With a Health Care Organization and Their Impacts on Health Quality Measurements: Protocol for a Retrospective Cohort Study," *J. Med. Internet Res.*, vol. 7, no. 11, 2018.

[387] A. Benis, R. B. Barkan, T. Sela, and N. Harel, "Communication Behavior Changes Between Patients With Diabetes and Healthcare Providers Over 9

Years: Retrospective Cohort Study," *J. Med. Internet Res.*, vol. 22, no. 8, p. e17186, 2020.

[388] A. M. Ali, M. D. Loeffler, P. Aylin, and A. Bottle, "Predictors of 30-Day Readmission After Total Knee Arthroplasty: Analysis of 566,323 Procedures in the United Kingdom," *J. Arthroplasty*, vol. 34, no. 2, pp. 242–248, 2019.

[389] P. Espinosa, R. J. Weiss, O. Robertsson, and J. Karrholm, "Sequence of 305,996 total hip and knee arthroplasties in patients undergoing operations on more than 1 joint," *Acta Orthop.*, vol. 90, no. 5, pp. 450–454, 2019.

[390] S. Sykes and H. Verbeek, "Creating a petri net using ProM 6.8 for conformance testing against an event log," *ProM Forum*, 2019. [Online]. Available: https://www.win.tue.nl/promforum/discussion/1306/creating-a-petri-net-using-prom-6-8-for-conformance-testing-against-an-event-log#latest. [Accessed: 10-Dec-2020].

[391] A. Adriansyah, "Aligning observed and modeled behavior," Technische Universiteit Eindhoven, 2014.

[392] National Institute for Health and Care Excellence, "Arthroscopic knee washout, with or without debridement, for the treatment of osteoarthritis," 2021. [Online]. Available: https://www.nice.org.uk/guidance/ipg230/chapter/4-Further-NICE-recommendations-on-the-treatment-of-osteoarthritis. [Accessed: 19-Jan-2021].

[393] B. C. Werner, M. Tyrrell Burrus, W. M. Novicoff, and J. A. Browne, "Total Knee Arthroplasty Within Six Months After Knee Arthroscopy Is Associated With Increased Postoperative Complications," *J. Arthroplasty*, vol. 30, no. 8, pp. 1313–1316, 2015.

[394] Q. Liu *et al.*, "The influence of prior arthroscopy on outcomes of primary total lower extremity arthroplasty: A systematic review and meta-analysis," *Int. J. Surg.*, vol. 98, p. 106218, 2022.

[395] A. Gu *et al.*, "Prior Knee Arthroscopy Is Associated With Increased Risk of Revision After Total Knee Arthroplasty," *J Arthroplast.*, vol. 35, no. 1, pp. 100–104, 2020.

[396] F. Emamjome, R. Andrews, and A. H. M. ter Hofstede, "A Case Study Lens on Process Mining in Practice.," in *On the Move to Meaningful Internet Systems: OTM 2019 Conferences. OTM 2019. Lecture Notes in Computer Science, vol 11877.*, 2019.

[397] K. Diba, K. Batoulis, M. Weidlich, and M. Weske, “Extraction, correlation, and abstraction of event data for process mining,” *WIREs Data Min. Knowl. Discov.*, vol. 10, no. 3, p. e1346, 2019.

[398] A. Woodie, “MongoDB Users Discuss Their NoSQL Journeys,” *Datanami*, 2017. [Online]. Available: https://www.datanami.com/2017/06/26/mongodb-users-discuss-nosql-journeys/.

[399] Y. Wanga, L. Kungb, and T. A. Byrd, “Big data analytics: Understanding its capabilities and potential benefits for healthcare organizations.,” *Technol. Forecast. Soc. Chang.*, vol. 126, pp. 3–13, 2018.

[400] NHS Wales Informatics Service, “Health in Wales.” [Online]. Available: https://www.wales.nhs.uk/ourservices/directory/Hospitals. [Accessed: 08-Apr-2021].

[401] C. Favre and H. Völzer, “The Difficulty of Replacing an Inclusive OR-Join,” in *Business Process Management. BPM 2012. Lecture Notes in Computer Science, vol 7481*, 2012.

[402] Cawemo, “Business Process Modeling.” [Online]. Available: https://cawemo.com/. [Accessed: 10-Feb-2022].

[403] Diagrams.net, “Save file formats.” [Online]. Available: https://www.diagrams.net/doc/faq/save-file-formats. [Accessed: 10-Feb-2022].

[404] M. T. Hannan, D. T. Felson, and T. Pincus, “Analysis of the discordance between radiographic changes and knee pain in osteoarthritis of the knee,” *Rheumatology*, vol. 27, no. 6, pp. 1513–7, 2000.

[405] D. Yu, K. P. Jordan, and G. Peat, “Underrecording of osteoarthritis in United Kingdom primary care electronic health record data,” *Clin. Epidemiol.*, vol. 10, pp. 1195–1201, 2018.

[406] The Phoenix Partnership (TPP), “ResearchOne: Transforming Data into Knowledge.” [Online]. Available: http://www.researchone.org/. [Accessed: 08-Dec-2016].

[407] E. Herrett *et al.*, “Data Resource Profile: Clinical Practice Research Datalink (CPRD),” *Int. J. Epidemiol.*, vol. 44, no. 3, pp. 827–836, 2015.

[408] A. Herbert, L. Wijlaars, A. Zylbersztejn, D. Cromwell, and P. Hardelid, “Data Resource Profile: Hospital Episode Statistics Admitted Patient Care (HES APC),” *Int. J. Epidemiol.*, vol. 46, no. 4, p. 1093–1093i, 2017.

[409] E. Coiera, “The forgetting health system,” *Learn. Heal. Syst.*, vol. 1, no. 4, p. e10023, 2017.

[410] NHS UK, “A&E Attendances and Emergency Admissions,” 2022. [Online]. Available: https://www.england.nhs.uk/statistics/statistical-work-areas/ae-waiting-times-and-activity/. [Accessed: 20-Apr-2022].

[411] B. F. Arnold, D. R. Hogan, Colford, J. M., and A. E. Hubbard, “Simulation methods to estimate design power: an overview for applied research,” *BMC Med. Res. Methodol.*, vol. 11, no. 94, 2011.

Appendix A

ICD-9 to ICD-10 code mapping

ICD-9 code	ICD-9 description	ICD-10 code	ICD-10 description
276	Disorders of fluid electrolyte and acid-base balance	E87	Disorders of fluid, electrolyte and acid-base balance
285	Other and unspecified anemias	D64	Other anemias
410	Acute myocardial infarction	I21	Acute myocardial infarction
414	Chronic ischemic heart disease	I25	Chronic ischemic heart disease
428	Heart failure	I50	Heart failure
486	Pneumonia, organism unspecified	J18	Pneumonia, unspecified organism
438	Late effects of cerebrovascular disease	I69	Sequelae of cerebro-vascular disease
536	Disorders of function of stomach	K31	Diseases of stomach and duodenum
427	Cardiac dysrhythmias	I49	Cardiac arrhythmias
401	Essential hypertension	I10	Essential (primary) hypertension
710	Diffuse diseases of connective tissue	M35	Systemic involvement of connective tissue

Appendix B
Disease trajectory models following the early rules

B.1 Rule 3

Figure B.1 presents the CVD trajectory model where ICD-9 codes are rounded to level three and the sequence number value is added in seconds to the timestamp. The model is displayed showing the top four percent of all resulting diagnoses.

Figure B.1 CVD trajectory model following Rule 3

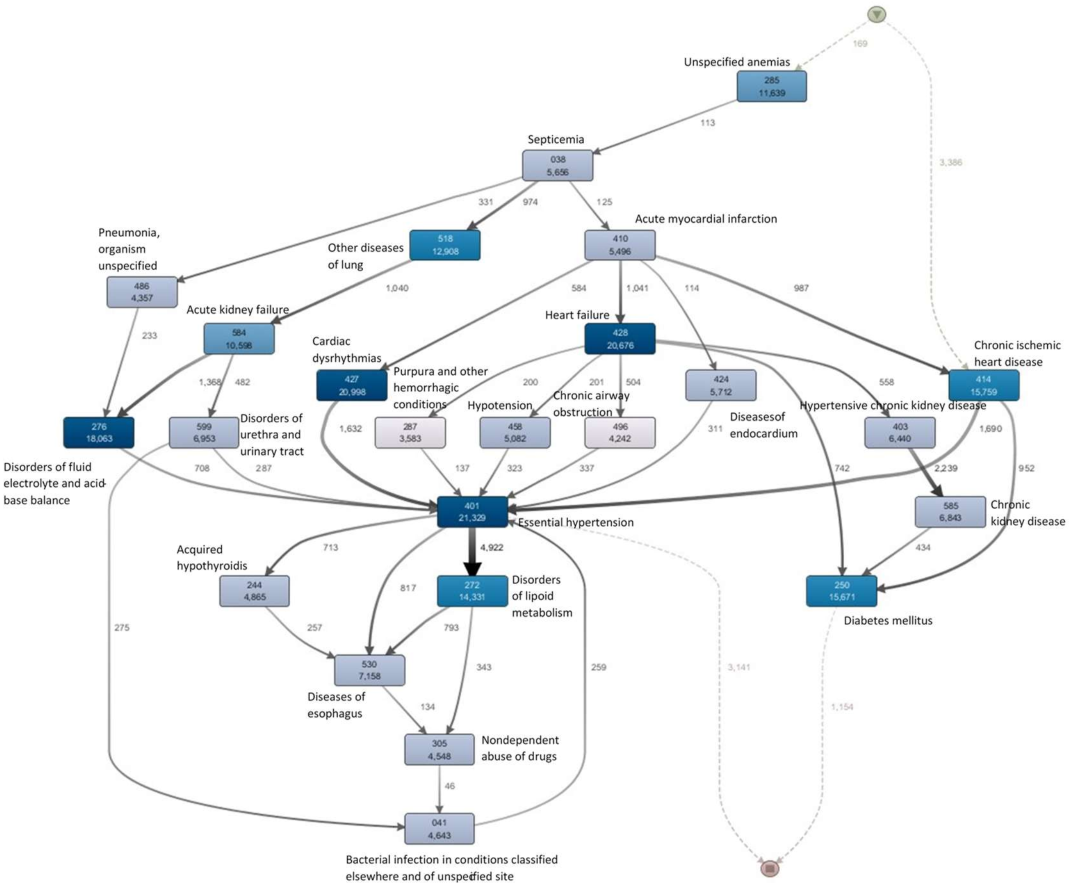
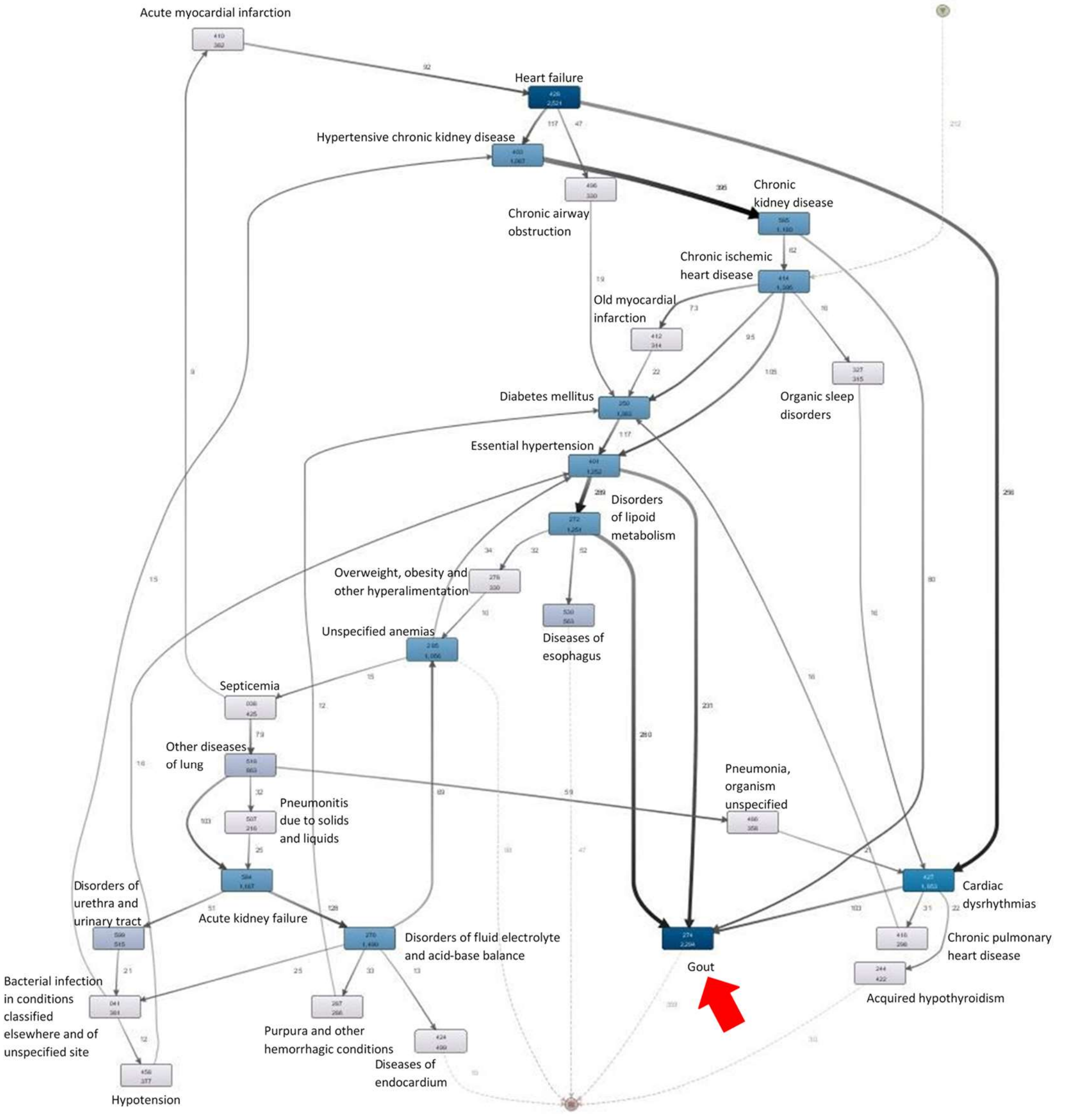


Figure B.2 shows the gout trajectories following Rule 3 displayed showing the top 6.6 percent of resulting diagnoses.

Figure B.2 Model for Rule 3 filtered on gout


Acute myocardial infarction
Heart failure
Hypertensive chronic kidney disease
Chronic kidney disease
Chronic airway obstruction
Chronic ischemic heart disease
Old myocardial infarction
Diabetes mellitus
Organic sleep disorders
Essential hypertension
Disorders of lipoid metabolism
Overweight, obesity and other hyperalimentation
Unspecified anemias
Diseases of esophagus
Septicemia
Other diseases of lung
Pneumonitis due to solids and liquids
Pneumonia, organism unspecified
Disorders of urethra and urinary tract
Acute kidney failure
Disorders of fluid electrolyte and acid-base balance
Gout
Cardiac dysrhythmias
Chronic pulmonary heart disease
Acquired hypothyroidism
Bacterial infection in conditions classified elsewhere and of unspecified site
Purpura and other hemorrhagic conditions
Diseases of endocardium
Hypotension

B.2 Rule 4

Figure B.3 presents the CVD trajectory model where ICD-9 codes rounded to level three with one second added to all comorbidities. This results in the primary diagnosis proceeding all comorbidities which have the same timestamp within a hospital admission. The model displays the top six percent of resulting diagnoses.

Figure B.3 CVD trajectory model following Rule 4

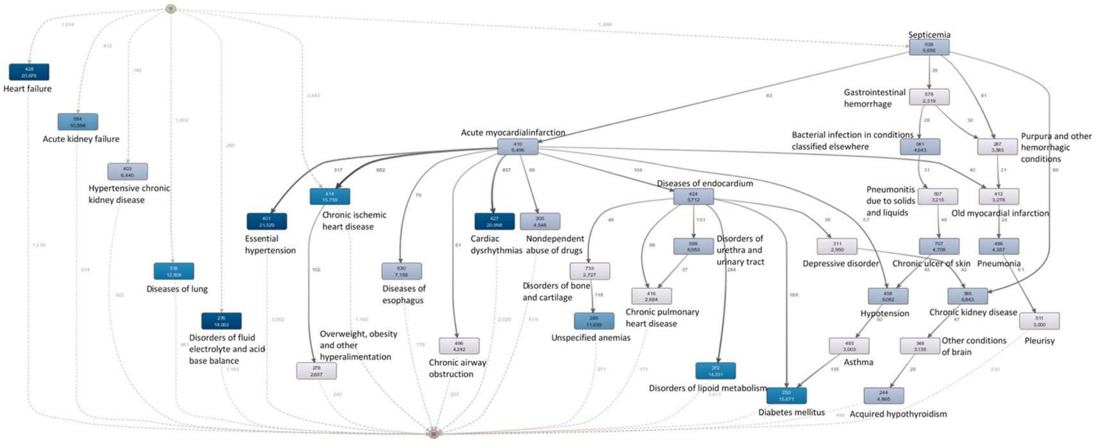
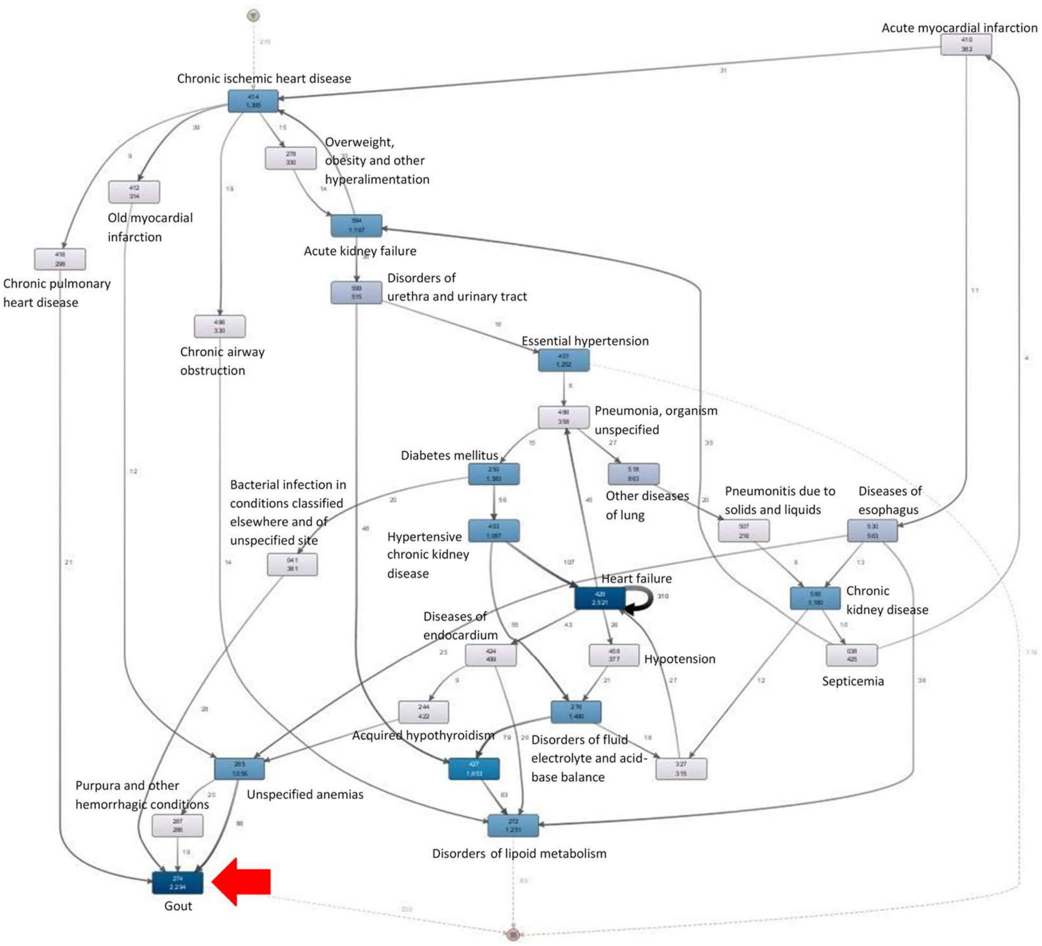


Figure B.4 shows the gout trajectories following Rule 4 and is displayed showing the top 6.6 percent of all resulting diagnoses.

Figure B.4 Model following Rule 4 filtered on gout


Acute myocardial infarction
Chronic ischemic heart disease
Overweight, obesity and other hyperalimentation
Old myocardial infarction
Acute kidney failure
Chronic pulmonary heart disease
Disorders of urethra and urinary tract
Chronic airway obstruction
Essential hypertension
Pneumonia, organism unspecified
Diabetes mellitus
Other diseases of lung
Pneumonitis due to solids and liquids
Diseases of esophagus
Bacterial infection in conditions classified elsewhere and of unspecified site
Hypertensive chronic kidney disease
Heart failure
Chronic kidney disease
Diseases of endocardium
Hypotension
Septicemia
Acquired hypothyroidism
Disorders of fluid electrolyte and acid-base balance
Purpura and other hemorrhagic conditions
Unspecified anemias
Disorders of lipoid metabolism
Gout

B.3 Rule 5

Figure B.5 presents the CVD trajectory model with level three ICD-9 codes, only primary diagnoses are included, therefore more than one hospital admission is required to form a trajectory. The model displays the top four percent of resulting diagnoses.

Figure B.5 CVD trajectory model following Rule 5

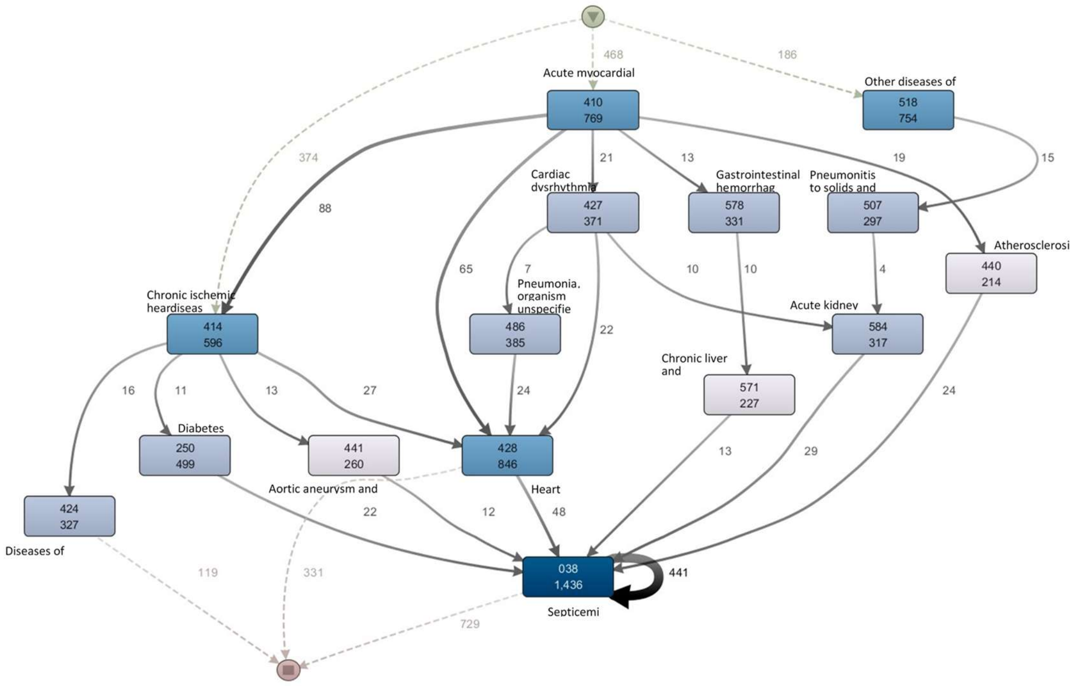
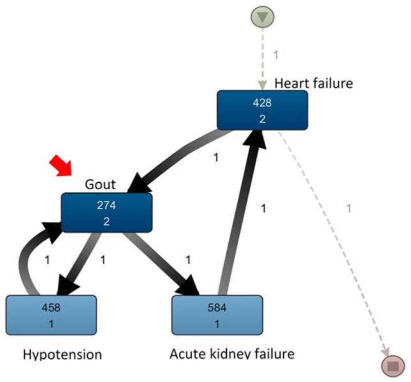


Figure B.6 shows the gout trajectories following Rule 5 and is displayed showing all resulting diagnoses.

Figure B.6 Model following Rule 5 filtered on gout



B.4 Rule 6

Figure B.7 presents the CVD trajectory model with level three ICD-9 codes, only primary diagnoses are included with repeating diagnoses excluded from future admissions, even when seen as a comorbidity. The model displays the top four percent of resulting diagnoses.

Figure B.7 CVD trajectory model following Rule 6

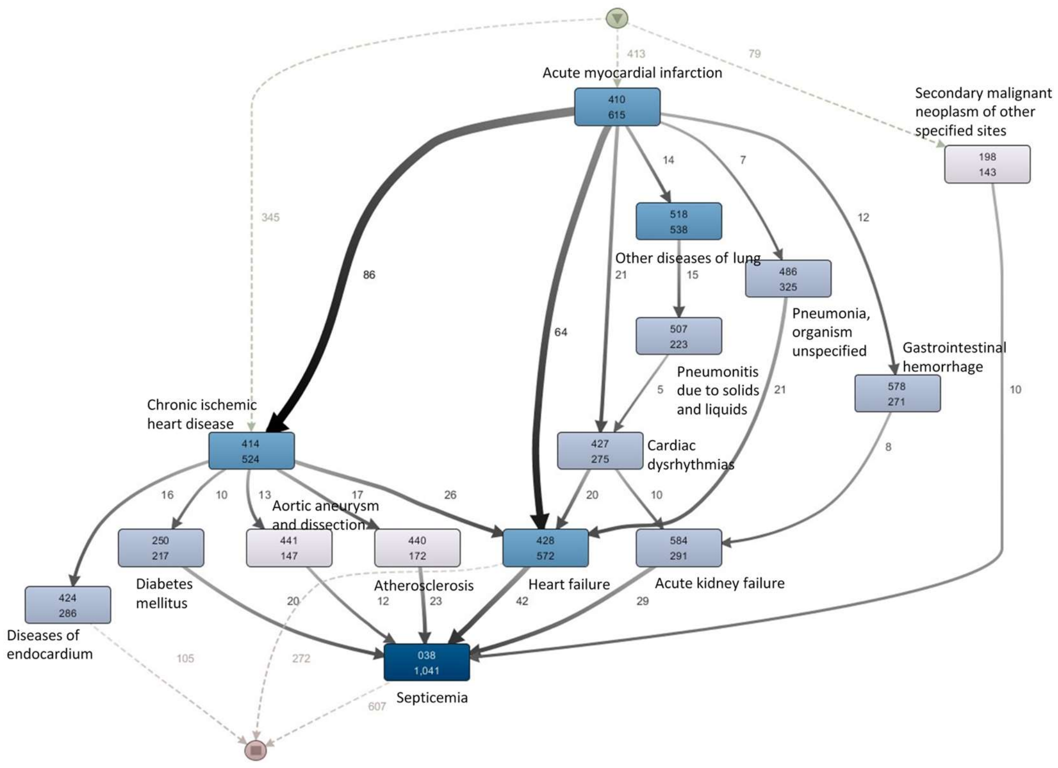
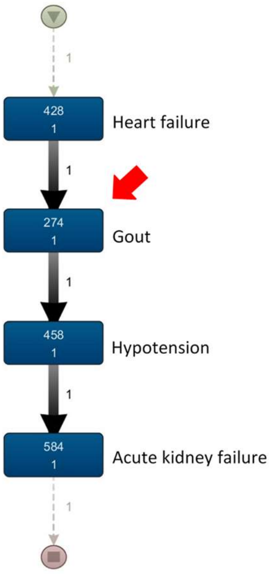


Figure B.8 shows the gout trajectory model following Rule 6 and shows all resulting diagnoses

Figure B.8 Model following Rule 6 filtered on gout



B.5 Rule 8

Figure B.9 presents the CVD trajectory model with level three ICD-9 codes and the sequence number value is added in seconds to the timestamp, repeating diagnoses are excluded. The model displays the top four percent of resulting diagnoses.

Figure B.9 CVD trajectory model following Rule 8

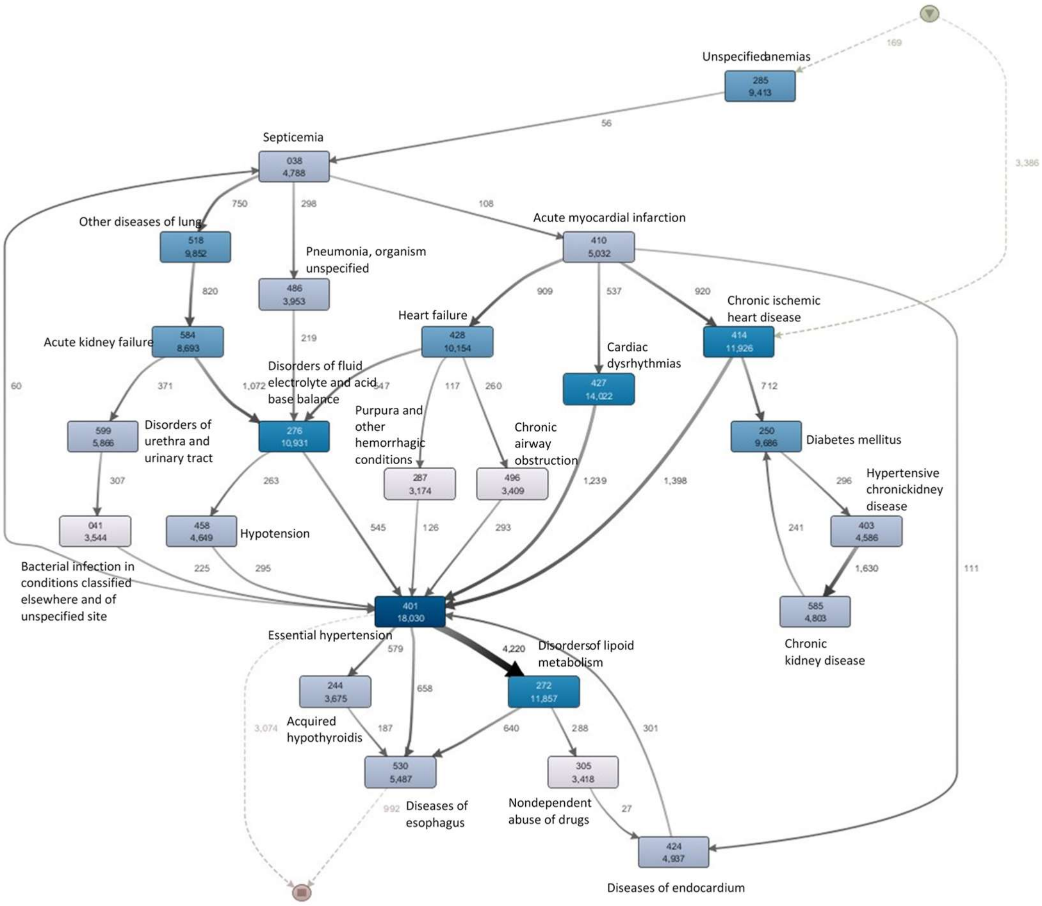
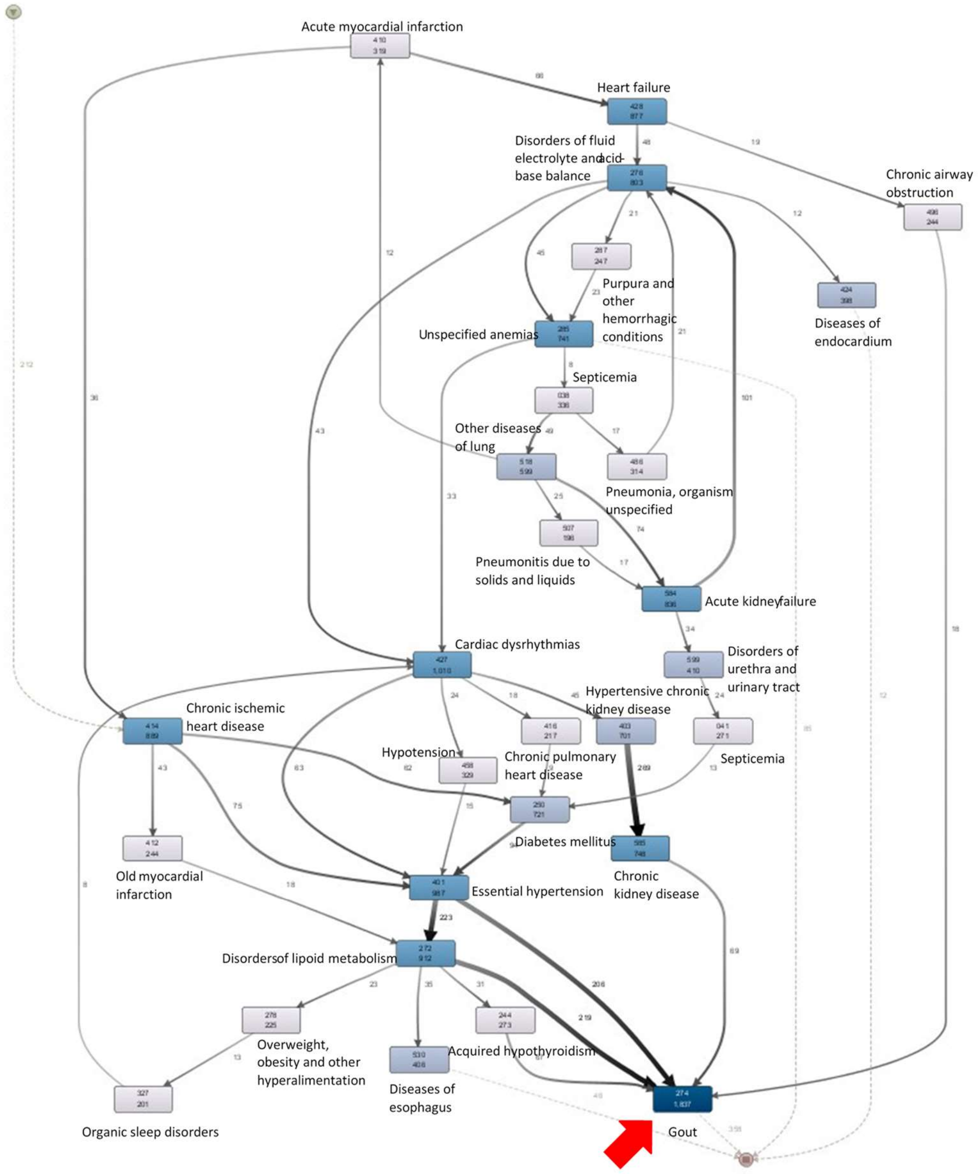


Figure B.10 shows the gout trajectory model following Rule 8 and is displayed showing to top 6.6 percent of diagnoses.

Figure B.10 Model following Rule 8 filtered on gout



Appendix C

MyPathway Extract, Transform and Load Issues Register

#	Date	File	Issue description	Imp	Resolution/comments
1	23-Jul	EQ	Need to match patients in extract to patient's questionnaire responses in Spreadsheet (SS).	H	Yes fid's are the same.
2	23-Jul	EQ	Need to match referrals in extract to referral questionnaire responses in SS. Need refId adding to completed questionnaire. Need to discuss if episode id's are needed and how will link to referrals?	H	Change to add referral id and Triage Decision into SS. Match on date in Python.
3	11-Sep	1	Can this event be included "view sheffieldachesandpains.com/ resource"?	L	No
4	11-Sep	1	For the events "ortho speciality referral status open", "pain speciality referral status open", "physiotherapy speciality referral status open", and "rheumatology speciality referral status open" closed, rejected and discharged - How is speciality identified?	L	For these 4 specialties combine as 'referral status open' - can get the different specialties from either the clinic code or the triage decision
5	11-Sep	1	Assigned priorities to the events for inclusion in extract	M	done
6	11-Sep	1	Rows missing event names	H	fixed
7	11-Sep	1	Is document id the UUID	M	Both are unique identifiers for the document (one created for SQL, the other for Mongo)
8	11-Sep	1	Same event present with different name - IPAQ	M	Resolved
9	11-Sep	1	Multiple "invitation updated" events	L	Taken out
10	12-Sep	1	ADI confused by the 'source' column - what is expected in this?	L	Column not needed
11	12-Sep		Post code displaying incorrectly	H	Fixed
12	12-Sep	2	Need method of the invitation in the event name	M	done
13	12-Sep	2	Need "expired" status for invitations	M	done

14	12-Sep	2	Questionnaires don’t have a method of delivery in title of event, is it possible to have this? E.g. ‘Email EQ5D Questionnaire Sent’. Is this a sensible request, or are all questionnaires emailed? Would have thought some were by letter.	M	All questionnaires assigned through the app.
15	12-Sep	2	Are all resources allocated by email? Again what about by post (letter)?	M	All through app
16	12-Sep	2	Is it possible to have the pathway, clinic and UUID against all events?	M	done
17	12-Sep	2	"EQ-5D questionnaire Assigned", 'IPAQ questionnaire Assigned' and 'Assign IPAQ questionnaire'- No method. No pathway - how do I know which triage decision and referral UUID this relates to, is it by the UUID?	M	done
18	12-Sep	2	"eq5d questionnaire in progress" and "ipaq questionnaire in progress" - Are these just implicit?	L	Now code there to capture these.
19	12-Sep	2	"send inpatient appointment", "send outpatient appointment" and "send phone appointment" - rather than having these as separate events better to have the type of appt tagged to the event name.	M	done
20	12-Sep	2	No questionnaires are expiring	M	Questionnaires expiring (some) should be coming through in more recent patients.
21	13-Sep	3	Is there a way to link episodes of care for patients (for a single triage decision)? Can link the referral data by document id, appointment data by document id, but not different kinds of data within an episode of care, i.e. referral data to appointment data. If we add a referral_id column as well as the document_id column?	H	done
22	14-Sep	4	"Allocate https://adi.cachefly.net/WelcomeToMyPathway.mp4 resource" and other resources - no refId and duplicated.	L	done
23	14-Sep	4	"referral open" and "referral triage decision made" - Copy the document_id in the referral_doc_id column?	M	done

24	14-Sep	4	"Invitation created - sms" and email and accepted all have no refId or pathway.	M	done
25	14-Sep	4	"questionnaire EQ-5D complete" and "questionnaire IPAQ complete" - No referral id or pathway (TD).	H	IPAQs done. EQ5Ds still issue - 1,373 completed. Remainder of cases solved with inferencing.
26	14-Sep	4	triage appt statuses arrived, called, seen and departed - are these needed?	M	They are meaningless. If they are confusing matters can filter out the states for triage decision appointments.
27	14-Sep	4	110 - Is this clinic code correct? It has 11 added when it relates to the triage.	L	Yes, both can have these triage function codes for either triage appointments or outpatient appointments.
28	14-Sep	4	"questionnaire unspecified Assigned" - Why is the questionnaire type unspecified? No clinic code.	L	Taken out
29	14-Sep	4 to 13	No invitation reminders in extract. Reminder emails appear to be sent as a regular email invitations. Please can these be labelled as "invitation email reminder"?	M	Rename 2nd ‘invitation created’ to ‘invitation reminder’
30	15-Sep	4	No DNAs	M	I can see DNAs (event name 'appointment status did-not-attend')
31	18-Sep	6	EQ5D - the assigned date is after the completed date.	L	corrected
32	19-Sep	6	eq5d - 2. a fair amount have completed entries and no assigned	M	Initial EQ-5Ds are assigned outside of trigger engine, so are not picked up in 'trigger engine log' processing. Would it be a problem if reintroduced duplicate questionnaire assignments?
33	19-Sep	6	What are these events? "Assign 026d0110-9c04-4ff0-8fe8-078fc76a2821 resource"?	L	Extract code error, will correct.
34	20-Sep	7	> 1 questionnaire assigned with same docId. Should these be labelled as reminders?	M	No, bug fixed
35	21-Sep	7	Problem with EQ5Ds. Getting duplicate EQ5Ds, neither have the TD or the referral id. The duplicates happen when there has NOT been a completed EQ5D. Completed EQ5Ds never have assigned EQ5D with a referral id.	H	If EQ-5D not assigned through trigger engine, tricky to link to referral. Might have to link to referral based on time. shall speak to Nigel to understand why the appointment is not linked to a referral.
			10,553 assigned EQ5Ds, 8,030 have no RefId		Not recording data properly at source.
			7,359 of the ones with no RefId are unique.		Issue resolved.

36	24-Sep	8	Oswestry and Keele questionnaires have the same document id.	L	Fixed
37	26-Sep	9	"0" is present in many "...appointment status…" event	H	Available-to-book appointments with no referral id. Some have 'preadmit' numbers which can sometimes be used to link to referrals. 10 corrected the zero RefIds for 70 patients, but leaves 3,171 patients with no RefId.
	01-Oct				Appointments made before MyPathway. These have episodeId 1:1 with refId. Others due to S1 inserting "0" when an appt is made. Here an earlier referral will have had "S1" as part of eventName, use same refId as earlier referral. Issue solved with with inference code.
38	26-Sep	9	Remove any row with empty or zero referral id	M	done
39	26-Sep	9	Remove any row with 'triage appointment status arrived', 'triage appointment status called', 'triage appointment status seen', 'triage appointment status departed', 'referral closed', 'referral open' Simon is concerned why 'referral closed', 'referral open' are not needed.	H	'referral closed' and 'referral open' events back in.
40	28-Sep	10	Python: create 2 event logs: 1 for analysis patient journeys and one for analysing referrals.	H	Fixed in inference code.
			See pseudo code for spec.		
41	28-Sep	10	For event log keyed on pid and refId	H	done
			All above plus:		
			Delete PATIENTS (all records) with null RefIds where title = “%questionnaire%”. Give count of deleted.		
			Delete PATIENTS (all records) with “0” RefIds. Give count of deleted.		
			Create a field where pid and refId are joined.		
42	28-Sep	10a	Problems reading full extract (2,401,935 rows) into Python. Line 408954 contained 2 records in 1.	H	Cut from |535|… and pasted to a new row (next) adding the Fid from the row below.
43	28-Sep	10a	Problems reading full extract into Python. Line 408966 (original file) only contained 4 pipes.	H	Removed row

44	28-Sep	10a	Problems reading full extract into Python. Line 851085 (after changes above) contained 2 records in 1.	H	Cut from second postcode … and pasted to a new row (next) adding the Fid from the row below (38987).
45	30-Sep	10a	last 3 records have incorrect fid	L	Need to check this has been resolved in full extract 13. Check and manually edit with new extract.
46	30-Sep	10a	From the log10a_PatientRefs file manually deleted 35,964 rows of type map resources or Welcome to MP resources where there was no refId, couldn't create a Case Key.	H	No solution, will put a note against the process maps that most maps and "welcome to MP" events are not present.
47	30-Sep	10a	In event logs some titles don't have new titles.	H	Changes made
48	30-Sep	10a	Why in the log10a_PatientRefs file is there only 14 "triage decision made" rows?	M	done
49	01-Oct	10a	These events should either be changed or not in extract as are meaningless: " appointment status available-to-book", " appointment status booked", "Outpatient appointment status", "departed", "inpatient appointment status", "outpatient appointment status", "outpatient appointment status undefined", "pre-admission appointment status", "questionnaire undefined complete", "triage appointment status", "triage appointment status arrived", "triage appointment status available-to-book", "triage appointment status called", "triage appointment status departed", "triage appointment status seen", "waiting-list appointment status".	M	done
50	01-Oct	10a	Create a new column in the extract file for patients that have only had 1 referral.	H	Added column containing flag.
51	01-Oct	10a	Add pre-admit number/episodeId column. So episodes of care can be tied to referrals when the triage decision has happened before MP.	H	done
52	02-May	10a	Reminder email appears to be sent as a regular email invitation.	M	done
53	02-May	10a	Check that the "outpatient appointment status seen, called, departed, waiting" are happening on the actual day of the appointment.	M	done
54	02-May	10a	Is confirmation email missing from events?		No

55	05-Nov	13	Patient xxxxx has only 1 referral id, with 2 'Referral open' events the 2nd is after a 'Referral discharged' event - is this ok?	H	Ok for referral to go from discharged to open with same number.
56	13-Nov	13	Do triage ddddE type preadmit nums sometimes turn into S1-type preadmit nums when they become real appointments?	H	A ddddE type PAN is often a triage PAN which can become a different ddddE type appointment PAN or an S1-type appointment PAN.
57	13-Nov	13	Regarding the rule where any events before the first ‘referral open’ event should be disregarded. What if the first event is ‘triage appt status booked’? or 'referral triage decision made'?	H	Should include with 'referral open' - 'triage appointment status booked', 'referral triage decision made' and 'outpatient appointment status available-to-book'.
58	13-Nov	13	There are 1,683 events with 0 RefId that have a preadmit_num like '1000000*1' these do not have any referral or triage events - what are they?	H	Ancient patients, none MP patients so discount.
59	14-Nov	13	How many cases have 1 preadmid_num assigned to >1 RefId?	H	3
60	27-Nov		How would you include the EQ VAS or wouldn’t you?	M	Do not use.
61	27-Nov	13	Should d EQ5Ds be identified by them being sent 125 days from the 1st appt booked or the 1st appt attended?	H	Code rule is from O/P appt booked date. It should be from the 1st attended. Will issue change request.
62	29-Nov	13	The code to split the EQ5Ds into baseline etc. is not working correctly.	H	Code not implemented.
63	29-Nov	13	Check names for phone appt's. There are phone appt attended and phone appt departed	M	renamed
64	03-Jan	13	If the patient was referred before MP or due to some other reason there may be no referral events in the data to link to.	H	Events before 15/05/2017 should be deleted.
65	03-Jan	13	Sometimes patients were referred then rejected, this can happen multiple times, therefore events without referral ids must link back to the correct referral id.	H	done
66	03-Jan	13	“waiting-list appointment status” events are present in the data. Rename “waiting-list appointment status” to ‘waitList appt booked’ for new event name.	H	done

67	03-Jan	13	‘inpatient appointment status admitted’ events are not linked with the same document id to ‘waiting-list appointment status (booked)’ events and should be.	H	done
68	03-Jan	13	There are multiple patient admitted and patient discharged events with the same document id.	H	done
69	10-Dec	13	Not recognising some of the EQ5D types.	H	Extract has pipes as separators.
					Solved.
70	10-Jan	14	Often EQ5Ds are issued at same time as the patient departed and the referral discharged. Which type should it be?	H	Discharge EQ5D
71	18-Jan	15	Given the information in the attached spreadsheet how close can we get to a diagnosis?	H	Bug in Lorenzo where Sheffield not getting Diagnosis Codes.
72	18-Jan	15	No baseline/triage decision EQ5Ds assigned.	H	TD/baseline EQ5Ds not assigned through rules - assigned through portal action. Only allocate baseline EQ5Ds outside of rules.
			Also sometimes getting a pathway and referral_doc_id in the EQ5D event and sometimes not. Some EQ5D assigned events don’t have rule names.		
73	21-Jan	15	can I confirm that:	H	No reminders being sent out for Baseline EQ5Ds, just for invitation. 1st appointment and discharge EQ5Ds should be being sent out.
			EQ5D reminders are not sent out		
			Only 2 triggers are connected with EQ5Ds which are:		
			a. sth-pathway-first-appointment-attended		
			b. sth-pathway-discharge		
			No EQ5D reminders?		
74	24-Jan	16	Is the best indicator of diagnosis triage decision?	M	See email on 11/03/19.
75	28-Feb	18	rule_names are sometimes being set to 'sth-pathway-first-appointment-attended' when they are not the 1st appt for the referral. Code is labelling them 'Pre-treatment EQ5D sent'.	H	fixed
76	01-Mar	18	Problem with matching of doc Ids for waiting-list and inpatient appointments.	L	Known issue but as not using docId will leave it.

77	30-Jul	19	In February extract some referral ids were present, but set to zero in July extract.	H	Can infer a refId if PW (Sys1) referrals.
78	23-Jul	19	2 triggers added to live system that are not used in code to identify EQ-5D types. These are rule_name = ‘sth-pathway-discharge-physio-send-eq5d’ and ‘sth-pathway-eq5d-for-triage-decision-entered’.	H	code added

Appendix D

MyPathway data cleansing and processing pseudo-code

Changes up to and including number 12 have been implemented by Dr Thamer Ba Dhafari in Python. Changes after number 12 have been implemented by Dr Mark Dixon in Java. The numbering matches the comment numbering in the code.

2. Delete events where title = "inpatient appointment status".
3. Delete events where title = "outpatient appointment status" or "Outpatient appointment status".
4. Delete events where title = "outpatient appointment status undefined".
5. Delete rows where NO Fid.
6. Delete "Allocate https://adi.cachefly.net/WelcomeToMyPathway.mp4 resource" events. (These should not be present in the extract)
8. Delete triage events:
    - "triage appointment status arrived"
    - "triage appointment status called"
    - "triage appointment status seen"
    - "triage appointment status departed"

8a. Delete duplicate AnkSpond Resource events.

```
IF title = ("Allocate https://nass.co.uk/ resource" OR "Allocate https://nass.co.uk/nass/en/about-as/living-well-with-as/ resource" OR "Allocate https://www.nass.co.uk/about-as/about-as/ resource"
    IF FID, eventDateTime, title, _title, and docId are duplicated (*docId must NOT be NULL, _title may be NULL)
        IF title NOT = "questionnaire EQ-5D Assigned"
            Delete the duplicate row WITHOUT the "preadmit_number"
```

9. Create new EventNames for all titles and insert into "_title" column.

11. Remove "T" and replace with a space from middle of timestamp. Remove .000Z from end of timestamp and ensure it includes seconds.

12. Delete all non–MyPathway patients (FID)

A MP patient has 1 of the following events:
"Invitation accepted – email"
OR
Any completed questionnaire event (see 23a for list)

Special rule - Whenever checking for a referral:
Check 1st for a 'referral open' event
In absence of "referral open" event check for "triage appointment status booked" event. In the absence of "referral open" and "triage appointment status booked" event check for "referral triage decision made" event
Any 1 of these 3 events can be a referral open event (ONLY USE ONE)

Special rule - when re-dating events (using the "appointment_date" instead of the "date" value. After replacing the date:

- Check event is still >= 15/05/2017

- Check there is a referral event before it
- If either of the above are false then delete the re-dated event
- Always re-sort the events on “date” for that patient

14. Delete the following test patients ('101294', '101299', '49772', '83581', '00000', '49805', '101297', '10721')

16. Delete all events belonging to a referral that has preadmit_nums in the format of dddddd*dd. The number of digit before and after the ‘*’ vary therefore easier to delete any that are not in the following formats:

   a. starts with ‘S1-‘
   b. ‘undefined’
   c. starts with 4 digits followed directly with ‘E’ (*ddddE*)
   d. is null

17. Rename the following events if an “Allocate %(M/m)ap%” event or an https://publicdocuments.sth.nhs.uk/pil627.pdf event within 180 seconds of it (either direction):

   - “outpatient appointment status booked” to “O/P appt-booked & map sent”
   - “waiting-list appointment status booked” to “Wait-list appt-booked & map sent”
   - “waiting-list appointment status” to “Wait-list appt-booked & map sent”
   - “telephone appointment status booked” to “phone appt booked & map sent”

   Delete map event (1 or 2 map events associated with appointment).

17b. Delete map events left over from 17 including https://publicdocuments.sth.nhs.uk/pil627.pdf.

19. If a patient is admitted (‘inpatient appointment status admitted’) **AND** discharged (‘inpatient appointment status discharged’) OR cancelled (‘inpatient appointment status cancelled’)), THEN either admitted, discharged or cancelled again for the same document id delete any admitted, discharged or cancelled events after the first set. They will all have the same document id.

20. Inference rules for ‘0’ Referral Id’s

```
    A. For a patient
IF preadmit_num starts with ‘S1-‘
        IF earlier ‘S1-‘ WHERE CCode was same
                Copy referral id AND pathway from CLOSEST previous ‘S1-‘ appt event with same CCode
                Copy referral id AND pathway to subsequent events WHERE referral id = (‘0’ OR NULL) AND preadmit_num is the same (as event being inferred)
        ELSEIF no earlier ‘S1-‘ WHERE CCode was same
                IF event = ('outpatient appointment status booked' OR 'waiting-list appointment status booked' OR 'telephone appointment status booked' OR 'outpatient appointment status departed' OR 'outpatient appointment status cancelled' OR 'outpatient appointment status did-not-attend' OR 'inpatient appointment status discharged' OR 'waiting-list appointment status cancelled' OR 'telephone appointment status cancelled' OR ‘O/P appt attended’ OR ‘O/P appt-booked & map sent’ OR ‘Wait-list appt-booked & map sent’ OR ‘ phone appt booked & map sent’) AND (AND a previous ‘ATB appt’ EXISTS WHERE Triage Decision = ‘Clinic PW*’) //where ‘*’ is a wildcard.
                //Start working backwards chronologically from the zero referral id event (so closest first)
                        IF NO booked appt EXISTS for the ‘ATB appt’ (matching the ATB and booked appt on refId)
                                IF the ‘ATB appt’ has a RefId that is not null or ‘0’ (if it has carry on looking)
                                        Copy RefId AND pathway from ‘ATB appt’
                                        Copy RefId AND pathway to subsequent events WHERE RefId = (‘0’ OR NULL) AND preadmit_num is the same
                ELSEIF event = ('outpatient appointment status booked' OR 'waiting-list appointment status booked' OR 'telephone appointment status booked' OR 'outpatient appointment status departed' OR 'outpatient appointment status cancelled' OR 'outpatient appointment status did-not-attend' OR 'inpatient appointment status discharged' OR 'waiting-list appointment status cancelled' OR 'telephone appointment status cancelled' OR ‘O/P appt attended’ OR ‘O/P appt-booked & map sent’ OR ‘Wait-list appt-booked & map sent’ OR ‘ phone appt booked & map sent’) AND (NO previous ‘ATB appt’ EXISTS WHERE Triage Decision = ‘Clinic PW…’)
                        Copy referral id AND pathway from CLOSEST previous ‘triage appointment status booked’ event
                        Copy referral id AND pathway to subsequent events WHERE referral id = (‘0’ OR NULL) AND preadmit_num is the same (as event being inferred)

    B. For a patient
IF preadmit_num = ‘undefined’
        IF an earlier ‘S1-‘ event WHERE CCode was the same
                Copy the RefId AND pathway from the closest ‘S1-‘ event
                Copy that RefId AND pathway to subsequent events WHERE RefId = (‘0’ OR NULL) and preadmit_num = ‘undefined’ AND docId = docId

    C. For a patient
IF preadmit_num starts with 4 digits followed directly with ‘E’ (nnnnE)
        IF only 1 previous non-rejected referral (matched on RefId)
                Copy referral id AND pathway from the ‘referral triage decision made’ event associated with the referral (matched on RefIf) OR if this event does not exist, from the ‘triage appointment status booked’ event, OR if this does not exist, from the ‘referral open’ event
        Copy that referral id AND pathway to subsequent events WHERE RefId = (‘0’ OR NULL) AND preadmit_num is the same
    ELSEIF > 1 previous non-rejected referral
        IF earlier non-rejected referral EXISTS WITHOUT an ('outpatient appointment status booked' OR 'waiting-list appointment status booked' OR 'telephone appointment status booked' OR 'outpatient appointment status departed' OR 'outpatient appointment status cancelled' OR 'outpatient appointment status did-not-attend' OR 'inpatient appointment status discharged' OR 'waiting-list appointment status cancelled' OR 'telephone appointment status cancelled' OR ‘O/P appt attended’ OR ‘O/P appt-booked & map sent’ OR ‘Wait-list appt-booked & map sent’ OR ‘ phone appt booked & map sent’)  event
            Copy referral id AND pathway FROM the CLOSEST (to the event being inferred) ‘referral triage decision made’ event associated with the referral (matched on RefIf) OR if this event does not exist, from the ‘triage appointment status booked’ event, OR if this does not exist, from the ‘referral open’ event
            Copy that referral id AND pathway to subsequent events WHERE RefId = (‘0’ OR NULL) AND preadmit_num IS SAME
        ELSEIF ALL earlier non-rejected referrals HAVE appointments
            Leave
20a.    IF RefId = ‘0’
        FOR a patient
        IF preadmit_number NOT = ‘undefined’, ‘0’ or NULL
            IF same preadmit_number exists (after or before row with ‘0’ refId – start working from the ‘0’ event into the future for that patient, then if none found from the ‘0’ event to the beginning for that patient)
                Copy referral id AND pathway to row with ‘0’ refId
13. Fix RefIds for questionnaires

    a. Delete duplicate “questionnaire EQ-5D Assigned” events

        IF FID, title and docId are duplicated (all 3 must be present, none of these can be null)
            IF title = “questionnaire EQ-5D Assigned”
                Delete duplicate WHERE “questionnaire_name” = ‘EQ-5D’
                Keep all values for the remaining row including “rule_name”

    a1. IF “questionnaire EQ-5D Assigned” events RefId IS NULL
        IF ‘preadmit_number’ exists for “questionnaire EQ-5D Assigned” event
            Copy RefId, pathway and IF rule_name IS NULL (for “questionnaire EQ-5D Assigned” event) copy rule_name from closest event for that patient WHERE ‘preadmit_number’ matches the “questionnaire EQ-5D Assigned” event with the NULL RefId

    a2. IF “questionnaire EQ-5D Assigned” events RefId IS NULL
        IF “questionnaire IPAQ Assigned” event exist for patient within 60 seconds in either direction of the “questionnaire EQ-5D Assigned” event
            Copy RefId, pathway and IF rule_name IS NULL rule_name from “questionnaire IPAQ Assigned” event

    a2a. IF “questionnaire EQ-5D Assigned” events RefId IS NULL
        IF “referral triage decision made” exists before this event
            IF “referral triage decision made” to “questionnaire EQ-5D Assigned” <= 60 seconds
                Copy the RefId from the “questionnaire EQ-5D Assigned” event

    a3. IF “questionnaire EQ-5D Assigned” events RefId IS NULL
        IF “questionnaire IPAQ expired” events RefId IS NULL
```

Copy RefId and pathway from "questionnaire IPAQ assigned" event with the same docId

14a. Replace 'date' timestamps for 'inpatient appointment status admitted' events with the 'appointment_date' value.

18. Delete events for a PATIENT with the same document id where title = "outpatient appointment status arrived" OR "outpatient appointment status called" OR "outpatient appointment status seen" OR "outpatient appointment status waiting" OR "outpatient appointment status departed" and replace with "O/P appt attended". Take the timestamp from the column 'appt date' of the last deleted event instead of the events timestamp (this will often be the day before). This means none of the above events will exist only one "O/P appt attended".

18a. Replace 'date' timestamps for 'telephone appointment status departed' events with 'appointment date' value.

15. Delete any events before 15/05/2017.

Delete any events for this patient (same FID) with same readmit_number OR refId (not including '0', NULL or 'undefined') as the event(s) just deleted, so long as event is NOT a 'referral open' or 'triage appointment status booked' or 'referral triage decision made' event. If event after 15/5/17 is one of these 3 do not delete event or any events with same preadmit_number or refId after it.

15a. Delete events before the 1st 'referral open', 'triage appointment status booked' OR 'referral triage decision made'.

Delete any events for this patient (same FID) with same preadmit_number OR refId (not including '0', NULL or 'undefined') as the event(s) just deleted.

21. 'inpatient appointment status admitted' events are not linked with the same document id to 'waiting-list appointment status (booked)' events and they should be. Therefore;

```
IF event = 'inpatient appointment status admitted' AND an earlier 'waiting-list   appointment
status booked' event exists
    WHERE patient id = patient id
    AND 'appointment date' (NOT TIME) = 'appointment date'
    AND referral id = referral id
    Replace the 'waiting-list appointment status booked' events' document id with the 'inpatient
    appointment status admitted' events' document id
    _title = 'wait-list appt booked'

ELSEIF event = 'inpatient appointment status admitted' AND an earlier 'Wait-list appt booked
& map sent' event exists
    WHERE patient id = patient id
    AND 'appointment date' (NOT TIME) = 'appointment date'
    AND referral id = referral id
    Replace the 'Wait-list appt booked & map sent' events' document id with the 'inpatient
    appointment status admitted' events' document id
    _title = 'Wait-list appt booked & map sent'
```

23a. Fill in referral ids for ALL completed and reminder questionnaires types –

(questionnaire BASDAI Assigned : questionnaire BASDAI complete : Reminder to Complete BASDAI questionnaire by email
questionnaire BASFI Assigned : questionnaire BASFI complete : Reminder to Complete BASFI questionnaire by email
questionnaire EQ-5D Assigned : questionnaire EQ-5D complete : Reminder to Complete EQ-5D questionnaire by email

questionnaire fft Assigned : questionnaire fft complete : Reminder to Complete fft questionnaire by email
questionnaire IPAQ Assigned : questionnaire IPAQ complete : Reminder to Complete IPAQ questionnaire by email
Assign SthKeele questionnaire : questionnaire SthKeele complete : Reminder to Complete SthKeele questionnaire by email
Assign SthOswestry questionnaire : questionnaire SthOswestry complete : Reminder to Complete SthOswestry questionnaire by email
Assign leeds_hip questionnaire : questionnaire leeds_hip complete : Reminder to Complete leeds_hip questionnaire by email
Assign leeds_knee questionnaire : questionnaire leeds_knee complete : Reminder to Complete leeds_knee questionnaire by email
questionnaire OxfordHip Assigned : questionnaire OxfordHip complete : Reminder to Complete OxfordHip questionnaire by email
Assign OxfordHip questionnaire : questionnaire OxfordHip complete : Reminder to Complete OxfordHip questionnaire by email
questionnaire SthPain Assigned : questionnaire SthPain complete : Reminder to Complete SthPain questionnaire by email)

```
    FOR ALL completed questionnaires (see list above)
    IF RefId IS NULL
        GET RefId, pathway and preadmit_num of assigned questionnaire (matching on
        document id)
        INSERT into completed questionnaire row

    Fill in referral id, pathway and preadmit_num for related questionnaire reminders:
    FOR ALL questionnaire reminders (see list above)
    IF RefId IS NULL
        GET RefId, pathway and preadmit_num of assigned questionnaire (matching on
        document id)
        INSERT into the reminder questionnaire event

    Fill in referral id, pathway and preadmit_num for questionnaire expired:
    FOR ALL “questionnaire IPAQ expired”
    IF RefId IS NULL
        GET RefId, pathway and preadmit_num of “questionnaire IPAQ Assigned”
        (matching on document id)
        INSERT into the “questionnaire IPAQ expired” event
```

23b. Any “questionnaire % Assigned”, “Reminder to Complete % questionnaire by email”, “questionnaire % complete”, “questionnaire % scored” and “questionnaire % expired” events without a referral id should be deleted.

24. If after doing the above there are still patient events with 0 or NULL RefIds delete the events only, as these will belong to referrals pre-MyPathway. Write the patient event history to a ‘remaining_ZeroOrNull_RefId_Patients.csv’ file.

25. Invitation reminders are currently labelled the same as invitations. Need to separate:

```
IF ‘invitation created – sms’
    IF ‘invitation created – sms’ exist before (matched on PID and RefId)
        Rename 2nd ‘invitation created – sms’ to ‘invitation reminder – sms’ // this should
        be in both title columns //

IF ‘invitation created – email’
    IF ‘invitation created – email’ exist before (matched on PID and RefId)
        Rename 2nd ‘invitation created – email’ to ‘invitation reminder – email’ // this
        should be in both title columns
```

27b. First and follow-up outpatient appointments need separating:

```
        IF “O/P appt attended” is the first one for patient referral (refId)
                “O/P appt attended” = “New O/P appt attended”
        ELSE “O/P appt attended” = “Follow-up appt attended”
```

26. Separating EQ5D types FOR A PATIENT REFERRAL

Extract is not identifying all Baseline EQ5Ds, therefore any ‘questionnaire EQ-5D Assigned’ events without a rule_name are ‘Baseline EQ5D sent’ events.

Also assigned EQ5Ds that have a value in the ‘rule_name’ column need labelling appropriately in the ‘_title’ column. All completed questionnaires need a type assigning.

Adding an _Title for Assigned questionnaires:

```
IF title = ‘questionnaire EQ-5D Assigned’ AND rule_name IS NOT NULL
        IF rule_name = ‘sth-pathway-discharge’
                _title = ‘Discharge EQ5D sent’
        ELSEIF rule_name = ‘sth-pathway-discharge-physio-send-eq5d’
                _title = ‘PhysioDischarge EQ5D sent’
        ELSEIF rule_name = ‘sth-pathway-triage-decision-made’
                _title = ‘Baseline EQ5D sent’
        ELSEIF rule_name = ‘sth-pathway-eq5d-for-triage-decision-entered’
                _title = ‘Baseline EQ5D sent’
        ELSEIF rule_name = ‘sth-pathway-first-appointment-attended’
                IF _title.‘Follow-up appt attended’ EXISTS for the RefId BEFORE the event
                        _title = ‘EQ5D Q sent’
                ELSE
                        _title = ‘Pre-treatment EQ5D sent’

ELSEIF title = ‘questionnaire EQ-5D Assigned’ AND rule_name IS NULL
        IF ‘referral triage decision made’ to ‘questionnaire EQ-5D Assigned’ <= 60 seconds ago
        (this is not matched on RefId)
                _title = ‘Baseline EQ5D sent’
                Copy pathway AND referralId from ‘referral triage decision made’ event
                Rule_name = ‘sth- pathway-triage-decision-made’
```

Adding an _Title for EQ-5D first and second reminder questionnaires

```
IF title = ‘Reminder to Complete EQ-5D questionnaire by email’
        IF rule_name = ‘sth-questionnaire-send-eq5d-first-reminder’
                _title = ‘1st reminder ’ + _title of event WHERE documentId = documentId AND
                title = ‘questionnaire EQ-5D Assigned’
                Copy pathway AND referralId from the assigned to the completed
        ELSEIF rule_name = ‘sth-questionnaire-send-eq5d-second-reminder’
                _title = ‘2nd reminder ’ + _title of event WHERE documentId = documentId AND
                title = ‘questionnaire EQ-5D Assigned’
                Copy pathway AND referralId from the assigned to the completed
```

26b. All completed EQ5Ds need matching against assigned EQ5Ds by document Id and given same questionnaire type, pathway and Referral Id.

```
    In the title column
        FOR “questionnaire EQ-5D complete”
        FIND “questionnaire EQ-5D Assigned” matching on docId
   In the _title column
```

```
IF value in _title column for the “questionnaire EQ-5D Assigned” event = ‘EQ5D Q sent’
        Insert into the _title column of the “questionnaire EQ-5D complete” event ‘EQ5D
        Q completed’
ELSE
        Copy type (1st word) from “questionnaire EQ-5D Assigned” event (Baseline, Pre-
        treatment, PhysioDischarge or Discharge) To “questionnaire EQ-5D complete”
        event and ADD “ EQ5D completed” to end of the string (e.g. “Baseline EQ5D
        completed”, “Pre-treatment EQ5D completed, “PhysioDischarge EQ5D
        completed” or “Discharge EQ5D completed”)
```

Copy pathway and Referral Id from “questionnaire EQ-5D Assigned” row to “questionnaire EQ-5D complete” row

27. Calculate health outcome for each referral where at least two matched completed EQ5Ds exist. First must be a Baseline or a Pre-treatment and matched one either an EQ5D Q, PhysioDischarge or a Discharge.

```
FOR a patient referral
Select RID, date, Response_String
WHERE
        (> 0 ‘Baseline EQ5D completed’) AND (> 0 ‘EQ5D Q completed’ OR > 0 ‘Discharge
        EQ5D completed’ OR > 0 ‘PhysioDischarge EQ5D completed’)
        OR
        (> 0 ‘Pre-treatment EQ5D completed’) AND (> 0 ‘EQ5D Q completed’ > 0 ‘Discharge
        EQ5D completed’ OR > 0 ‘PhysioDischarge EQ5D completed’)
```

Remove any questionnaires WHERE Response_String format NOT= “0-4,0-4,0-4,0-4,0-4,0-100”.
Take only the first and last questionnaires for a referral.
Remove any matched questionnaires that are <= 21 days apart.

Calculate the health outcome

First completed EQ5D	Last completed EQ5D
“a1,a2,a3,a4,a5,aV”	“b1,b2,b3,b4,b5,bV”

```
IF
        (a1 < b1) AND (a2 <= b2 AND a3 <= b3 AND a4 <= b4 AND a5 <= b5)
        OR
        (a2 < b2) AND (a1 <= b1 AND b3 <= a3 AND b4 <= a4 AND b5 <= a5)
        OR
        (a2 < b2) AND (a1 <= b1 AND b3 <= a3 AND b4 <= a4 AND b5 <= a5)
        OR
        (a2 < b2) AND (a1 <= b1 AND b3 <= a3 AND b4 <= a4 AND b5 <= a5)
        OR
        (a2 < b2) AND (a1 <= b1 AND b3 <= a3 AND b4 <= a4 AND b5 <= a5)
THEN health_outcome = ‘Declined’

ELSEIF
        (a1 = b1 AND a2 = b2 AND a3 = b3 AND a4 = b4 AND a5 = b5)
THEN health_outcome = ‘No change’

ELSEIF
        (a1 > b1) AND (a2 >= b2 AND a3 >= b3 AND a4 >= b4 AND a5 >= b5)
        OR
        (a2 < b2) AND (a1 >= b1 AND b3 >= a3 AND b4 >= a4 AND b5 >= a5)
        OR
        (a2 < b2) AND (a1 >= b1 AND b3 >= a3 AND b4 >= a4 AND b5 >= a5)
        OR
        (a2 < b2) AND (a1 >= b1 AND b3 >= a3 AND b4 >= a4 AND b5 >= a5)
        OR
        (a2 < b2) AND (a1 >= b1 AND b3 >= a3 AND b4 >= a4 AND b5 >= a5)
THEN health_outcome = ‘Improved’

ELSE
        Health_outcome = ‘Mixed’
```

Insert Health_outcome value into last column of log file

Create 6 new columns in the log file ‘mobility’, ‘selfCare’, ‘usualActivities’, ‘painDiscomfort’, anxietyDepression’ and ‘eqVas’. Insert an integer from the string in the ‘questionnaire_answers’ column into each new column starting from the beginning of the string into the ‘mobility’ column.

27a. Delete ‘inpatient appointment status discharge’ events as timestamps are same as admitted.

28. Create output file containing everything, an error file with the zero and null referral id patients and the following output files:

a. All patient referrals
b. Patient referrals with improved health outcomes, e.g. file ‘cleaned_part1_dump19_<bodyPart>_<healthOutcome>.csv’ for:
   i. All body parts
   ii. Back
   iii. Foot
   iv. Elbow
   v. Hand/Wrist
   vi. Hip
   vii. Knee
   viii. Shoulder
c. Patient referrals with declined health outcomes
   i. All body parts
   ii. Back
   iii. Foot
   iv. Elbow
   v. Hand/Wrist
   vi. Hip
   vii. Knee
   viii. Shoulder
d. Patient referrals with no change health outcomes
   i. All body parts
   ii. Back
   iii. Foot
   iv. Elbow
   v. Hand/Wrist
   vi. Hip
   vii. Knee
   viii. Shoulder
e. Patient referrals with mixed health outcomes
   i. All body parts
   ii. Back
   iii. Foot
   iv. Elbow

v. Hand/Wrist
vi. Hip
vii. Knee
viii. Shoulder

Column headings for output files:

PatientID, Age, Gender, RefID, Date, pathway, _title, Q_title, questionnaire_answers, mobility, selfCare, usualActivities, painDiscomfort, anxietyDepression, eqVas.

Appendix E

MyPathway event log Extract, Transform and Load table

Title from extract	Eventlog title	T	L	Merged reason	T	L	Split reason	E	T	L	Deletion reason	Issue / Clean #
Allocate https://adi.cachefly.net/WelcomeToMyPathway.mp4 resource								✓			Welcome message sent for all patient referrals.	I22/C6
Allocate map (over 100 different URL titles)	O/P appt-booked & map sent	✓		Logic used to identify related events.						✓	Often entered days or weeks after they were booked so timestamp is invalid.	C17
Allocate map (over 100 different URL titles)	Wait-list appt-booked & map sent	✓		As above						✓	Physiotherapy patients do not have operations.	C17
Allocate map (over 100 different URL titles)	phone appt-booked & map sent	✓		As above						✓	No booking events used	C17
Assign QuickDASH questionnaire	Assign QuickDASH questionnaire								✓		Too few	
Assign SthKeele questionnaire	send SthKeele Q								✓		As above	
Assign SthOswestry questionnaire	send Oswestry Q								✓		As above	
inpatient appointment status								✓			Meaningless event.	I49/C2

inpatient appointment status admitted	I/P appt-admitted								✓		Physiotherapy patients do not have operations.	
inpatient appointment status discharge									✓		Meaningless, timestamp always same as admitted.	C27a
inpatient appointment status cancelled	I/P appt-cancelled									✓	Physiotherapy patients do not have operations.	
invitation accepted - email	invitation accepted-email									✓	On-boarding events - not of concern for health outcomes.	
invitation accepted - letter	invitation accepted-letter									✓	On-boarding event	
invitation accepted - sms	invitation accepted-sms									✓	As above	
invitation created - email	invitation sent-email				✓		Reminders not identified			✓	As above	C25
invitation created - email	invitation reminder - email				✓		As above			✓	As above	C25
invitation created - letter	invitation sent-letter									✓	As above	
invitation created - sms	invitation sent-sms				✓		Reminders not identified			✓	As above	C25
invitation created - sms	invitation reminder - sms				✓		As above			✓	As above	C25
outpatient appointment status								✓			As above	I49/C3
Outpatient appointment status								✓			As above	I49/C3
outpatient appointment status available-to-book	O/P appt-available-to-book											
outpatient appointment status booked	O/P appt-booked & map sent	✓		Logic used to identify related events.						✓	Retrospectively entered, timestamp invalid.	C17

outpatient appointment status arrived	O/P appt attended --> Follow-up O/P appt attended	✓		Unreliable timestamp	✓		Either first (New) or Follow-up					C18/27b
outpatient appointment status arrived	O/P appt attended --> New O/P appt attended	✓		As above	✓		As above					C18/27b
outpatient appointment status called	O/P appt attended --> Follow-up O/P appt attended	✓		As above	✓		As above					C18/27b
outpatient appointment status called	O/P appt attended --> New O/P appt attended	✓		As above	✓		As above					C18/27b
outpatient appointment status cancelled	O/P appt-cancelled											
outpatient appointment status departed	O/P appt attended --> Follow-up O/P appt attended	✓		Unreliable timestamp	✓		Either first (New) or Follow-up					C18/27b
outpatient appointment status departed	O/P appt attended --> New O/P appt attended	✓		As above	✓		As above					C18/27b
outpatient appointment status did-not-attend	O/P appt DNA											
outpatient appointment status seen	O/P appt attended --> Follow-up O/P appt attended	✓		Unreliable timestamp	✓		Either first (New) or Follow-up					C18/27b
outpatient appointment status seen	O/P appt attended --> New O/P appt attended	✓		As above	✓		As above					C18/27b
outpatient appointment status undefined								✓			Meaningless event	I49/C4

outpatient appointment status waiting	O/P appt attended --> Follow-up O/P appt attended	✓		Unreliable timestamp	✓		Either first (New) or Follow-up					C18/27b
outpatient appointment status waiting	O/P appt attended --> New O/P appt attended	✓		As above	✓		As above					C18/27b
questionnaire BASDAI complete	completed BASDAI Q									✓	Too few	
questionnaire BASFI complete	completed BASFI Q									✓	As above	
questionnaire EQ-5D Assigned	PhysioDischarge EQ5D sent				✓		1 of 5 types			✓	Not live.	C26
questionnaire EQ-5D Assigned	Baseline EQ5D sent				✓		As above			✓	not a predictor for health outcome.	C26
questionnaire EQ-5D Assigned	Discharge EQ5D sent				✓		As above			✓	As above	C26
questionnaire EQ-5D Assigned	EQ5D Q sent				✓		As above			✓	As above	C26
questionnaire EQ-5D Assigned	Pre-treatment EQ5D sent				✓		As above			✓	As above	C26
questionnaire EQ-5D complete	PhysioDischarge EQ5D completed				✓		As above			✓	As above	C26b
questionnaire EQ-5D complete	Baseline EQ5D completed				✓		As above			✓	As above	C26b
questionnaire EQ-5D complete	Discharge EQ5D completed				✓		As above			✓	As above	C26b
questionnaire EQ-5D complete	EQ5D Q completed				✓		As above			✓	As above	C26b
questionnaire EQ-5D complete	Pre-treatment EQ5D completed				✓		As above			✓	As above	C26b

questionnaire fft complete	completed fft Q									✓	Too few	
questionnaire IPAQ Assigned	send IPAQ Q									✓	As above	
questionnaire IPAQ complete	completed IPAQ Q									✓	As above	
questionnaire IPAQ expired	IPAQ Q expired									✓	As above	
questionnaire leeds_hip complete	completed LeedsHip Q									✓	As above	
questionnaire leeds_knee complete	completed LeedsKnee Q									✓	As above	
questionnaire OxfordHip complete	completed OxHip Q									✓	As above	
questionnaire PEM Assigned	PEM Q sent									✓	As above	
questionnaire QuickDASH complete	questionnaire QuickDASH complete									✓	As above	
questionnaire SthKeele complete	completed SthKeele Q									✓	As above	
questionnaire SthOswestry complete	completed SthOswestry Q									✓	As above	
referral closed	referral closed									✓	Automatic system event with no business meaning.	
referral discharged	referral discharged									✓	Can mean discharge for patient referral or referral to different speciality, depending on clinician.	
referral open	referral open											

Reminder to attend appointment by email	Appt reminder-email								✓		Not predictor for health outcome.	
Reminder to Complete EQ-5D questionnaire by email	1st reminder PhysioDischarge EQ5D sent				✓		1 of 5 types. There are 1st and 2nd reminders.		✓		As above	C26
Reminder to Complete EQ-5D questionnaire by email	1st reminder Baseline EQ5D sent				✓		As above		✓		As above	C26
Reminder to Complete EQ-5D questionnaire by email	1st reminder Discharge EQ5D sent				✓		As above		✓		As above	C26
Reminder to Complete EQ-5D questionnaire by email	1st reminder EQ5D Q sent				✓		As above		✓		As above	C26
Reminder to Complete EQ-5D questionnaire by email	1st reminder Pre-treatment EQ5D sent				✓		As above		✓		As above	C26
Reminder to Complete EQ-5D questionnaire by email	2nd reminder PhysioDischarge EQ5D sent				✓		As above		✓		As above	C26
Reminder to Complete EQ-5D questionnaire by email	2nd reminder Baseline EQ5D sent				✓		As above		✓		As above	C26
Reminder to Complete EQ-5D questionnaire by email	2nd reminder Discharge EQ5D sent				✓		As above		✓		As above	C26
Reminder to Complete EQ-5D questionnaire by email	2nd reminder EQ5D Q sent				✓		As above		✓		As above	C26
Reminder to Complete EQ-5D questionnaire by email	2nd reminder Pre-treatment EQ5D sent				✓		As above		✓		As above	C26

Reminder to Complete leeds_hip questionnaire by email	Leeds_hip Q email reminder								✓		Too few	
Reminder to Complete leeds_knee questionnaire by email	LeedsKnee Q email reminder								✓		As above	
Reminder to Complete OxfordHip questionnaire by email	OxHip Q email reminder								✓		As above	
Reminder to Complete PEM questionnaire by email	PEM Q email reminder								✓		As above	
Reminder to Complete QuickDASH questionnaire by email	Reminder to Complete QuickDASH questionnaire by email								✓		As above	
Reminder to Complete SthKeele questionnaire by email	SthKeele Q email reminder								✓		As above	
telephone appointment status arrived	phone appt	✓		Automated event meaning phone app								
telephone appointment status booked	phone appt-booked	✓								✓	Retrospectively entered, timestamp invalid.	C17
telephone appointment status booked	phone appt-booked & map sent	✓		Logic used to identify related events.						✓	As above	C17
telephone appointment status cancelled	phone appt-cancelled									✓	Too few	
telephone appointment status departed	phone appt	✓		Automated event meaning phone app								

telephone appointment status did-not-attend	phone appt-DNA									✓	Too few	
telephone appointment status seen	phone appt	✓		Automated event meaning phone app								
triage appointment status arrived available-to-book								✓			Automated, meaningless	I49/C8
triage appointment status								✓			As above	I49/C8
triage appointment status called								✓			As above	I49/C8
triage appointment status departed								✓			As above	I49/C8
triage appointment status seen								✓			As above	I49/C8
triage appointment status booked	triage appt booked									✓	Not predictor for health outcome.	
triage appointment status cancelled	traiage appt cancelled									✓	As above	
referral triage decision made	triage decision made									✓	Triggers at the same time as ‘O/P appt-available-to-book’.	
waiting-list appointment status	Wait-list appt booked	✓		Type error in source system (should include booked)								I66/C21
Wait-list appt booked	Wait-list appt booked	✓		Merge with event above						✓	Physiotherapy patients do not have operations.	I66/C21
Wait-list appt booked	Wait-list appt-booked & map sent	✓		Logic used to identify related events.						✓	As above	C17/C21

waiting-list appointment status cancelled	wait-list appt cancelled									✓	As above	1

[1] For the purpose of this table the data cleansing and data processing has been combined and labelled as ‘T’ for transform. In some cases events were initially merged or split, before being later deleted after analysis of the results for specific experiments.

Appendix F
OPCS-4 Knee pain surgery codes

Laterality:

Z941 Bilateral operation

Z943 Left sided operation

Z942 Right sided operation

Qualifiers:

Z844 Patellofemoral

Z846 Knee joint

F.1 Arthroscopy codes

W701 Open total excision of semilunar cartilage

W702 Open excision of semilunar cartilage NEC

W703 Open repair of semilunar cartilage

W708 Other specified open operations on semilunar cartilage

W709 Unspecified open operations on semilunar cartilageW821

W822 Endoscopic resection of semilunar cartilage NEC

W823 Endoscopic repair of semilunar cartilage

W828 Other specified therapeutic endoscopic operations on semilunar cartilage

W829 Unspecified therapeutic endoscopic operations on semilunar cartilage

W851 Endoscopic removal of loose body from knee joint

W852 Endoscopic irrigation of knee joint

W853 Endoscopic autologous chondrocyte implantation of knee joint

W858 Other specified therapeutic endoscopic operations on cavity of knee joint

W859 Unspecified therapeutic endoscopic operations on cavity of knee joint

W871 Diagnostic endoscopic examination of knee joint and biopsy of lesion of knee joint

W878 Other specified diagnostic endoscopic examination of knee joint

W879 Unspecified diagnostic endoscopic examination of knee joint

Codes needing qualifiers:

W711 Open drilling articular cartilage

W833 Endoscopic shaving artic cartilage

F.2 Primary TKR codes

W401 Primary total prosthetic replacement of knee joint using cement

W408 Other specified total prosthetic replacement of knee joint using cement

W409 Unspecified total prosthetic replacement of knee joint using cement

W411 Primary total prosthetic replacement of knee joint not using cement

W418 Other specified total prosthetic replacement of knee joint not using cement

W419 Unspecified total prosthetic replacement of knee joint not using cement

W421 Primary total prosthetic replacement of knee joint NEC

W428 Other specified other total prosthetic replacement of knee joint NEC

W429 Unspecified other total prosthetic replacement of knee joint

Codes needing qualifiers:

W541 Primary prosthetic replacement of articulation of bone NEC

W548 Other specified other prosthetic replacement of articulation of other bone

W549 Unspecified other prosthetic replacement of articulation of other bone

W581 Primary resurfacing arthroplasty of joint

F.3 TKR revision codes

W400 Conversion from previous cemented total prosthetic replacement of knee joint

W402 Conversion to total prosthetic replacement of knee joint using cement

W403 Revision of total prosthetic replacement of knee joint using cement

W404 Revision of one component of total prosthetic replacement of knee joint using cement

W410 Conversion from previous uncemented total prosthetic replacement of knee joint

W412 Conversion to total prosthetic replacement of knee joint not using cement

W413 Revision uncemented total knee replacement

W414 Revision of one component of total prosthetic replacement of knee joint not using cement

W420 Conversion from previous total prosthetic replacement of knee joint NEC

W422 Conversion to total prosthetic replacement of knee joint NEC

W423 Revision of total prosthetic replacement of knee joint NEC

W425 Revision of one component of total prosthetic replacement of knee joint NEC

W426 Arthrolysis of total prosthetic replacement of knee joint

Codes needing qualifiers:

W543 Revision of prosthetic replacement of articulation of bone NEC

W580 Conversion from previous resurfacing arthroplasty of joint

W582 Revision of resurfacing arthroplasty

Attention to TKR codes:

W424 Attention to total prosthetic replacement of knee joint NEC

W913/Q Other manipulation of joint, Manipulation of prosthetic joint nec

Appendix G

GQFI Table

Goal	Purpose	To Assess		
	Issue	the face validity and the clinical plausibility of the		
	Process	knee pain surgery pathway model and results		
	Viewpoint	from a clinician's viewpoint		

Question	PM Feature	Indicators	Values	Domain expert comments
Can a knee pain surgery reference model be defined for patients diagnosed with knee pain using process mining techniques?	Process discovery (Celonis)	Number of initial surgery event types	4	This is in accordance with what the experts would expect.
		Initial surgery event types	Right and left: Arthroscopy, Primary TKR	This is in accordance with what the experts would expect.
		Number of final surgery event types	4	This is in accordance with what the experts would expect. Some very uncommon surgeries were deliberately omitted and should be specified as a limitation e.g. high tibial osteotomy.
		Final surgery event types	Right and left: TKR Revision, Attention to	As above
		Number of connections	31	The level of connections displayed was set related to frequencies. This was understandable by the clinicians.
		Direction and frequency of connections between the surgery events		Considered a useful way to validate the sequence and frequency of surgery events for a patient. Visualising these connections helped to validate relevant patient events that were in accordance with what the experts would expect. E.g. multiple arthroscopies; and primary TKR surgery happened before TKR revision surgery but not in the opposite direction.
	Conformance checking (ProM)	Matching rate *(2nd iteration)*	100.0%	Helped to validate the model using the data. Though may have been an inefficient use of time for small number of violations.
		Matching rate *(1st iteration)*	98.5%	This was expected to be high because the number of surgeries was limited.
		Number of violations *(1st iteration)*	507	These violations were due to missing or incorrect data and this was an expected low number due to the limited number of surgeries on the model.
	Process enhancement (DISCO and manual)	Additional low-level surgeries *(identified via Process variant analysis, then manual process after exporting list of patients and investigating base data)*	3	This helped to enhance the model by broadening the capture of relevant events. The three codes were correctly identified.
Does the behaviour (sequence of surgery events) in the real-life SAIL data conform to the knee pain surgery reference model?	Conformance checking (ProM)	Matching rate *(1st iteration)*	98.5%	Given the confidence in the data (previous comments) and after viewing the data within the model using the fitness and variant visualisations along with viewing the low level data for the violations, these numbers appear correct.
		Number of violations *(1st iteration)*	507	As above
	Process variant analysis (ProM)	Number of violations *(1st iteration)* - Due to wrong events in the data - Due to missing events in the data	507 304 203	Some degree of violations, due to miscoding, is expected as it is routinely collected data. However, the low percentage of violations supports the model and the data.

Can useful healthcare statistics be generated from the SAIL data for patients with knee pain using process mining techniques?	Event log inspection (ProM)	Min/mean/max number of surgeries per patient	1/1/16.	This is in accordance with what the experts would expect.
		Min/mean/max number of surgery types per patient	1/1/6.	This is in accordance with what the experts would expect.
	Process variant analysis (ProM)	Number of variants	567	As expected, a high level of variability exists for this process within the data.
		Percentage of patients per variant (where percentage of patients > 1%) Right Primary TKR Left Primary TKR Right Arthroscopy Left Arthroscopy Left Primary TKR -> Right Primary TKR Right Primary TKR -> Left Primary TKR Right Arthroscopy -> Right Arthroscopy Left Arthroscopy -> Left Arthroscopy Right Arthroscopy -> Right Primary TKR Left Arthroscopy -> Right Arthroscopy Left Arthroscopy -> Left Primary TKR	17.40% 15.42% 12.51% 10.58% 6.33% 5.59% 3.19% 2.76% 2.41% 1.78% 1.81%	A cut-off point of 1% is considered reasonable. This is useful information to understand service utilisation and provide data to triage services for planning purposes. Using these figures, clinicians can get overall prevalence of knee joint replacement within their population in order to advise patients on their chances of having a joint replacement post arthroscopy, etc. The slightly higher percentage of right-sided surgeries was in accordance with what the experts would expect, as this is usually the dominant side for the patient. The limitation of duration of follow-up has previously been noted in Section 8.3.3.2.
		Outliers (Patient group size = 1 for the variant)	3.13%	There must be some patient variation and system variation that allows for some outliers. The frequency is as low as expected.
	Process discovery (Celonis)	For patients with single Primary TKR surgery Percentage with right TKR (n=2,053) And with >= 1 arthroscopy Or with >= 1 revision Percentage with left TKR (n=1,757) And with >= 1 arthroscopy Or with >= 1 revision	71% 19% 5% 69% 17% 4%	The close match (within 2%) of the results from the Swedish Registry data [354] provides support for the validity of this method. Overall this data is useful for clincal planning, as previously described. This is in accordance with what the experts would expect. The reason for fewer arthroscopies prior to a second knee replacement may be 1) the patient is younger at the time of the first joint replacement and therefore probably has less severe osteoarthritis, warranting an arthroscopy [393]. At the time of the second joint replacement the osteoarthritis may be too advanced; 2) the patient and surgeon may be aware that the arthroscopy was unsucessful the previous time and therefore reluctant to have a second; and 3) the NICE guidelines have changed to only allow surgeons to perform arthroscopies on patients under strict conditions [357].
		For patients with 2 Primary TKR surgeries Percentage with left TKR as second (n=845) And with >= 1 left arthroscopy Or with >= 1 right arthroscopy Percentage with right TKR as second (n=793) And with >= 1 right arthroscopy Or with >= 1 left arthroscopy	29% 13% 19% 31% 11% 17%	

Can useful healthcare statistics be generated from the SAIL data for patients with knee pain using process mining techniques? (continued)	Performance analysis (Celonis)	Time (median) between surgeries for patients with single TKR surgery Right arthroscopy -> right arthroscopy (n=55) Left arthroscopy -> left arthroscopy (n=45) First right arthroscopy to right TKR (n=391) First left arthroscopy to left TKR (n=295) Right TKR to first right revision (n=57) Left TKR to first left revision (n=45) Right TKR to first attention to right TKR (n=50) Left TKR to first attention to left TKR (n=33)	 818 d (2.2 yr) 1,293 d (3.5 yr) 771 d (2.1 yr) 695 d (1.9 yr) 893 d (2.4 yr) 834 d (2.3 yr) 229 d (0.6 yr) 184 d (0.5 yr)	These figures are all within the range of what the experts would expect, apart from the time between TKR and revision for patients with two TKRs. The figures for TKR to revision for patients having single TKR surgery are in agreement with figures published in the National Joint Registry 2020 [172].
		Time (median) between surgeries for patients with two TKRs For first TKR First right arthroscopy to right TKR (n=164) First left arthroscopy to left TKR (n=131) Right TKR to first right revision (n=21) Left TKR to first left revision (n=27) Right TKR to first attention to right TKR (n=12) Left TKR to first attention to left TKR (n=16) For second TKR Right TKR to left TKR (n=845) Left TKR to right TKR (n=793) First right arthroscopy to right TKR (n=88) First left arthroscopy to left TKR (n=113)	 625 d (1.7 yr) 605 d (1.7 yr) 2,221 d (6.1 yr) 1,753 d (4.8 yr) 112 d (0.3 yr) 1,038 d (2.8 yr) 764 d (2.1 yr) 769 d (2.1 yr) 1,161 d (3.2 yr) 1,389 d (3.8 yr)	A possible explanation for the duration between TKR to first revision being much longer for patients having surgery on both knees may be the following: Patients with bilateral OA that have had surgery on their first knee may develop problems but opt for the second knee surgery before addressing any problems with the first. After having both knees operated on, they may be reluctant to undergo a revision.
		Percentage of patients that undergo "unnecessary" arthroscopies (arthroscopies performed in close proximity to a TKR): - Percentage of knee replacement patients that have undergone primary TKR surgery with a previous arthroscopy on the same knee - Out of those patients that had an arthroscopy, the percentage that had their TKR within 12 months	5.2% 17% 5.2%	This is a way of identifying potentially inappropriate surgeries by identifying surgeries on the same knee that happened within close temporal proximity. The information can then be used to make potential improved efficiencies within the system. This is in accordance with what the expects would expect. Werner et al. [393] stated that between 2.2% and 10.2% of patients with osteoarthritis who have knee arthroscopy will undergo TKR surgery within one year of their arthroscopy. This is an important finding, though it is reasonably well known.
		(Does arthroscopy surgery before primary TKR surgery increase the risk of complications and further surgery?) - Revision rate for patients having an arthroscopy followed by a primary TKR, followed by a revision TKR compared to the revision rate for patients not first having an arthroscopy - Revision rate for patients having an arthroscopy followed by a primary TKR, followed by attention to the TKR compared to the revision rate for patients not first having an arthroscopy	Yes 0.046 compared to 0.020 0.030 compared to 0.015	There appears to be approximately double the prevelance of revision surgery in those who have had an arthroscopy prior to TKR. This warrants consideration by the field of orthopaedic surgery. The following two studies [394], [395] have identified that patients receiving knee arthroscopy before total knee replacement surgery are at a substantially increased risk of revision. This is an important finding, though it is reasonably well known.