Analysing the Impact of Changes in User Interface of e-Health Record Systems on Clinical Pathways using Process Mining

Angelina Prima Kurniati

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

The University of Leeds

School of Computing

December 2019

Intellectual Property and Publication Statement

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

Part of the work in the chapters of this thesis has appeared in jointly-authored publications as listed below:

Chapter 2

Kurniati, A.P., Johnson, O., Hogg, D. and Hall, G., 2016, October. Process Mining in Oncology: A Literature Review. In *IEEE International Conference on Information Communication and Management (ICICM), (pp. 291-297).*
The work in this paper was contributed and written by Kurniati, A.P.
Supervision, feedback, and general guidance were provided by Johnson, O., Hogg, D. and Hall, G.

Chapter 3

Johnson, O., Ba Dhafari, T., Kurniati, A.P., Fox, F., and Rojas, E., 2018, September. The ClearPath Method for Care Pathway Process Mining and Simulation. Lecture Notes in Business Information Processing (LNBIP).
The work in this paper was contributed and written by Johnson, O.
Kurniati, AP., was contributed and written on the 'Literature Review' and 'Challenges using EHR data for process mining care pathways' sections.

Chapter 4

1. **Kurniati, A.P., Johnson, O.A., Hogg, D. and Hall, G.**, 2017. Data Quality Issues with Using the MIMIC-III Data for Process Mining in Healthcare. In *Journal of Innovation in Health Informatics 2017.*

   The work in this poster was contributed and written by Kurniati, A.P. Supervision, feedback and guidance were provided by Johnson, O., Hogg, D. and Hall, G.

2. **Kurniati, A.P., Rojas, E., Johnson, O., Hogg, D. and Hall, G.**, 2018. The Assessment of Data Quality Issues for Process Mining in Healthcare using MIMIC-III, a Freely Available e-Health Record Database. In *Health Informatics Journal*.
   The work in this paper was contributed and written by Kurniati, A.P. Supervision, feedback, and guidance were provided by Rojas E., Johnson, O., Hogg, D. and Hall, G.
3. **Kurniati, A.P., Hall, G., Hogg, D. and Johnson, O.**, 2018, March. Process mining in Oncology using the MIMIC-III dataset. In Journal of Physics: Conference Series (Vol. 971, No. 1, p. 012008). IOP Publishing.
   The work in this paper was contributed and written by Kurniati, A.P. Supervision, feedback, and guidance were provided by Johnson, O., Hogg, D. and Hall, G.

Chapter 5

Kurniati, A.P., Hall, G., Hogg, D. and Johnson, O., 2018, November. Process Mining to Explore Variation in Chemotherapy Pathways for Breast Cancer Patients. The 2018 NCRI Cancer Conference (poster), the British Journal of Cancer Supplement, Springer Nature, Abstract no. 42 (vol.119 no.11 pp.16).
The work in this poster was contributed, written, and presented by Kurniati, A.P. Supervision, feedback, and general guidance were provided by Johnson, O., Hogg, D. and Hall, G.

Chapter 6

Kurniati, A.P., McInerney, C., Zucker, K., Hall, G., Hogg, D. and Johnson, O., 2019, September. A multi-level approach for identifying process change in cancer pathways. The 2019 Process Oriented Data Science for Health (PODS4H).
The work in this poster was contributed and written by Kurniati, A.P.
Feedback, general guidance, and supervision were provided by McInerney, C., Zucker, K., Johnson, O., Hogg, D. and Hall, G.

Acknowledgements

I would like to thank my supervisors Owen Johnson, Professor David Hogg, and Professor Geoff Hall, for their exceptional guidance and supervision during my study. This thesis would never be completed without their patience, encouragement, and constructive critiques.

This research has been carried out within a research team of healthcare process mining. This team has been led by Owen Johnson and has included Frank Fox, Amirah Alharbi, Nik Fatinah, Samantha Sykes, Samantha Crossfield, Guntur Kusuma, and me. My contributions, fully and explicitly indicated in the thesis, have been on process mining in oncology, specifically working on healthcare data from Leeds Teaching Hospitals NHS Trust (LTHT) and the process changes over time. The other members of the group have been working on other datasets on different focuses. Their contributions have been built through discussions during my PhD study. Special thanks to Eric Rojas, Thamer Ba Dhafari, Kieran Zucker, and Ciaran McInerney who have been generously sharing their knowledge, resources, and precious time with me.

I would also like to thank all members of the development team at the Informatics Department in the LTHT who provided access to the data, gave me unlimited resources and information to understand many aspects of the clinical and technical sides of the data.

Last but not least, I would like to thank my family and friends, both in Leeds and Indonesia. Support, understanding, and patience of Guntur, Gendis, and Panji are truly my strength during all ups and downs of my study. Endless appreciation is for my parents, Pak Priyanto and Bu Chris, and my younger siblings, Lora and Dryan, and also my parents in law, Papa Yoss and Mama Putu. This thesis is dedicated to you all.

Abstract

The provision of care in a hospital includes a series of activities that are often recorded in the electronic health record (EHR) systems. Analysing the data in these EHRs has the potential to support the understanding of care processes and exploring the opportunities for process improvement.

One of the emerging data analytics approaches for such analyses is process mining, and one critical challenge in working with EHR data is that processes might change over time. This thesis uses a process mining approach to detect process change over time and analyse the impact of those changes on the EHR data. The overall aim is to summarise the attributable change in the data due to process so that clinicians can better analyse the data.

Three datasets were used in this study to understand the variability of the EHR systems. The first dataset is a publicly available EHR data that was used for developing the methods and supporting the reproducibility of the research. The second dataset is a de-identified subset of the database of cancer patients from the Leeds Cancer Centre. The second dataset was used in the experiments to improve on the results of a previous study using the same dataset. The third dataset was the full Leeds Cancer Centre EHR database after more comprehensive ethics was approved. In the third dataset, experiments were done to analyse the impact of a known system change on clinical pathways and to explore process change over time without a known system change. All three datasets were analysed using process mining.

Process mining was shown to be useful for analysing clinical pathways and exploring process changes over time. It can be used to visualise the process before and after a known change. When the system change is unknown, process mining can be used to explore the process execution over time and identify the potential period where the system was changed. This thesis explores some aspects of the complex inter-relatedness of process and user interface (UI) of the EHR system.

Table of Contents

List of Tables

List of Figures

List of Terminology

activity : a well-defined step in a process

case : an individual experience of a process of interest

concept drift : the phenomenon that a process changes over time

conformance checking : checking if the process model discovered conform to the records in the event log, and vice versa

event : an activity related to a particular case

event log : a collection of events as input for process mining

fitness : the ability of a model to accurately reproduce the cases recorded in the log

generalisation : The ability of a process model to reproduce future behavior of the process

pathway/ path : a class of identical or broadly similar traces

precision : the ability of a model to precisely represent only behaviours recorded in the event log

process : a set of actions or steps taken in order to achieve a particular end

process discovery : a task in process mining to create a process model based on an event log

process mining : An approach to discover, monitor, and improve processes based on the event log automatically recorded in a computerised system

process model : an abstract model of the essence of a process which reflects to common pathways

sequence : a set of events and transitions for a specific case in a particular order

trace : a sequence of events for a case as recorded in the event log

transition : represent the link from one event to a following event to create a sequence event

variation : how and when a trace varies from the assumed process model

List of Abbreviations

A&E	:	Accident and Emergency
APN	:	Augmented Petri Nets
BIDMC	:	Beth Israel Deaconess Medical Centre
BPA-H	:	Business Process Analysis in Healthcare environments
BPIC	:	Business Process Intelligence Challenge
BPMI	:	Business Process Management Initiative
BPMN	:	Business Process Modelling Notation
bupaR	:	Business Process Analysis in R
CCG	:	Clinical Commissioning Group
CT	:	Computerised Tomography
CQC	:	Care Quality Commission
CRISP-DM	:	Cross-Industry Standard Process for Data Mining
CV	:	Philips CareVue
DISCO	:	Discovery tool for process mining
edeaR	:	Exploratory and Descriptive Event-data Analysis in R
E-R	:	Entity-Relationship
E& V	:	Initial encounter and external causes (ICD-9 codes)
ED	:	Emergency Department
EHR	:	Electronic Health Record
EPC	:	Event-driven Process Chain
ETL	:	Extract, Transform, Load
GP	:	General Practitioner
HHS	:	Health and Human Services (US)
HIPAA	:	Health Insurance Portability and Accountability Act (US)
HIS	:	Healthcare Information System
ICD	:	International Classification of Diseases (e.g ICD-9, ICD-10)
ICD-O	:	International Classification of Diseases for Oncology
ICICM	:	International Conference on Information Communication and Management
ICO	:	Information Commissioner’s Office
ICU	:	Intensive Care Units
iDHM	:	Interactive Data-Aware Heuristics Miner
IQR	:	Interquartile Range
IRAS	:	Integrated Research Application System
KTP	:	Knowledge Transfer Partnership
KTPId	:	Knowledge Transfer Partnership Identifier
LCR	:	Leeds Care Record
LHP	:	Leeds Health Pathways
LIDA	:	Leeds Institute for Data Analytics
LoS	:	Length of Stay
LTHT	:	Leeds Teaching Hospital NHS Trust
MDL	:	Minimal Description Length
MDT	:	Multidisciplinary Team
MIG	:	Medical Interoperability Gateway
MIMIC-III	:	Medical Information Mart for Intensive Care III
MIT	:	Massachusetts Institute of Technology
MRI	:	Magnetic Resonance Imaging
MV	:	iMDSoft MetaVision
NCI	:	National Cancer Institute
NHS	:	National Health System
NICE	:	National Institute for Health and Care Excellence
NIH	:	National Institutes of Health
NSCLC	:	non-small-cell lung carcinoma

OMG	Object Management Group
ONS	Office for National Statistics
PDM	Process Diagnostics Method
PHE	Public Health England
PHE NCRAS	Public Health England National Cancer Registration and Analysis Service
PID	Patient Identification
PM^2	Process Mining Project Methodology
PODS4H	Process-Oriented Data Science for Healthcare
PPM	Patient Pathway Manager
ProM	Process Mining framework
SCLC	small-cell lung carcinoma
SIGN	Scottish Intercollegiate Guidelines Network
SJUH	St James’s University Hospital (UK)
SNOMED	Systematised Nomenclature of Medicine
SPC	Statistical Process Control
TRIP	Turning Research into Practice
UI	User Interface
UK	United Kingdom
UML	Unified Modelling Language
UML AD	Unified Modelling Language Activity Diagram
UoL IRC	University of Leeds Integrated Research Campus
US	United States
WfMC	Workflow Management Coalition
WHO	World Health Organisation
WUN	World University Network
XES	eXtensible Event Stream
XML	eXtensible Markup Language

Chapter 1
Introduction

1.1 Overview

The provision of care in a hospital is delivered through healthcare processes. Healthcare processes can be described as a series of activities in the diagnosis, treatment, and follow-up of any disease aimed at improving patient health [1]. In modern healthcare, these activities are often recorded in a structured way using computerised information systems within the hospital. The information systems collect, store and manage data about these healthcare processes [2]. Analysing these data could be beneficial to build an understanding of healthcare processes, differences in processes, what results in the best outcome, and opportunities for improvement. As data volume grows, data analysis needs to improve to keep pace. This is especially important when considering that a large amount of data is collected during healthcare transactions every day. In the last few decades, many advancements in data analytics approaches have been introduced. Process mining is one of the emerging approaches that has the potential to offer new and interesting insights that help improve healthcare [3, 4].

The most common type of computerised information systems in a hospital is the Electronic Health Record (EHR) system [5, 6]. Clinicians use EHR systems to record the activities that they do in relation to the care they provide. Examples of activities in a treatment process within a hospital are screening, admission, investigation, pathological test, surgery, chemotherapy, and radiotherapy. The flow from one activity to the others for a patient in a clinical setting is known as a *care pathway* [7] or *clinical pathway* [8]. Analysis of the recorded data in the EHR system is potentially useful to assess the quality of the care pathways.

Process mining is an analytic approach to discover, monitor, and improve a process by analysing the data about that process [3, 4]. The input of process mining is an *event log*, which contains data about the process of interest. Process mining proposes a systematic approach to use that event log to create *process models*. A process model is a representation of the process from a specific perspective [4]. The discovered process model can then be used to check the conformance of the individual cases, and to improve the process. Process mining has been used in many case studies and various organisations, such as education [9], insurance [10], and healthcare [11]. In a healthcare setting, a care pathway is a process to be analysed with process mining.

One critical challenge in process mining is that the process might change over time [3]. Process mining is most useful when analysing a large volume of data, which is not easily analysed with manual approaches. Such data may be collected over a long period. A common approach in analysing these data in process mining is to assume that the process has not changed during the period of the study. In reality, processes are frequently changed over time. There are many reasons for that, for example, the organisation needs to align with a new procedure, the staff find an alternative way to do a process, a software upgrade, or the *User Interface (UI)* of the information system might need to be improved by introducing a new feature.

This thesis explores the opportunity for accessing EHR data and using a process mining approach to analyse care pathways, considering the possibility of process changes over time. The motivation of this thesis is to help make the healthcare process mining community more aware of the impact of process changes in process mining projects. The aims are: (1) to develop methods based on process mining to analyse the impact of changes in UI of EHR systems on clinical pathways, (2) to examine the applicability of process mining on different EHR datasets, and (3) to analyse EHR data to retrospectively examine the treatment process on Leeds patients diagnosed with cancer. This thesis focuses on the changes in the UI of EHR systems as a starting point of the analyses and discussed related aspects of system changes over time. Three datasets were used in this study to investigate the variability of EHR data. The challenges of this study need to draw on the insights from computer science and medical science.

1.2 Understanding the problem domain

To illustrate the range of challenges the thesis will explore, this section describes an example of a patient journey in a hospital. This example is based on a fictitious person entirely based on the discussions with clinical experts. The example is explored from the health service, process mining and information system perspectives.

1.2.1 An illustrative example

An illustrative example is provided by following the journey of Jane, a patient in a hospital. Jane was referred by her General Practitioner (GP) on an urgent basis to a gynaecologist in a hospital. The gynaecologist met her in an outpatient consultation. After checking on the detailed history and examination, the gynaecologist requested a range of tests (investigations) including an ultrasound scan, blood test, and hysteroscopy. Hysteroscopy is a procedure where a fibre optic camera is used to directly visualise the internal anatomy of the womb and obtain biopsy samples if required. The results of the investigations were then discussed in a multi-disciplinary team (MDT) review. The MDT review is held to get a consensus view of the gynaecologist, oncologist, radiologist, pathologist, and other healthcare professionals. Jane's pathology sample confirmed a diagnosis of cancer. The outcome of the MDT review was a recommendation to proceed to surgical treatment.

A few days later, Jane brought back to the hospital to meet with a gynaecological oncologist. During the appointment, the test results and cancer diagnosis were explained to Jane. They discussed the recommended course of surgery as the first and definitive treatment. Jane agreed and was formally consented for surgical treatment. Jane attended the hospital two weeks later where she was admitted for her surgical procedure. The operation successfully removed all of the visible tumours and she was discharged awaiting results of the surgical histology. At the follow up appointment, the surgeon met Jane explaining that the operation was successful and that the pathology showed healthy tissue and clear margins from her surgical resection. Jane was then followed up with annual outpatient visits for five years after which she was discharged with no signs of recurrent disease.

Jane's case is used as an example to illustrate one possible pathway for cancer treatment in a typical NHS hospital providing specialised cancer services. This example will now be further discussed from the health service perspective, the process mining perspective, and the information system perspective.

1.2.2 Health service perspective

The important stakeholders from the health service perspective are the patients, the clinicians, and the health service managers. Most obviously, the patient is the main stakeholder, along with their family who cares about the sequence of patient treatment. Jane’s journey can be seen as one example of an individual care pathway from the patient perspective. For Jane, the patient, it is important to know the details of each activity, who will deliver the treatment, when and how long it will last, and what are the results of the treatment. It is upsetting for the patient to be diagnosed with cancer, so the more detail given and the clearer their understanding about the individual pathway the better. When the gynaecologist did some tests, for example, Jane needs to know what will be tested, how will it be done and what to expect. When Jane and the oncologist discuss the course of treatment, Jane needs to know her options and risks of each option. The gynaecologist needs to explain this based on his or her understanding of previous care pathways that he or she is aware of.

Another important stakeholder is the clinician who executes the activities within the process. This refers to a clinician with specific expertise, but also as part of a Multidisciplinary Team (MDT) within a hospital. The clinician is interested in the treatment given to a patient under their care. In Jane’s example, when the gynaecologist sees Jane in a consultation, the gynaecologist needs to know what has been done to the patient, what her current condition is, and what should be done as the next step in the pathway. When the clinicians meet in an MDT meeting, they need to discuss Jane’s care pathway from different perspectives. Process mining can be used to analyse durations, variants, and pathways from different perspectives needed by clinicians.

Within a hospital, a health service manager is also a stakeholder who cares about patient treatments in the hospital. The health service manager is responsible for directing, coordinating, and administering medical and non-medical resources, facilities and services. Health service managers would be interested in Jane’s journey to see how the medical and non-medical resources within the hospital provide health services for patients. Health service managers need to plan, direct, and coordinate health services. From Jane’s example, the health service managers manage the medical resources, facilities, and services in the oncology unit, the pathology unit, and the surgery unit to support Jane’s treatment. They need to plan the services, analyse how the services have been used, how clinicians work from time to time, and what can be done to improve overall performance. They might need to see the sequence of Jane’s treatment and to know if the sequence conforms to the standard guidelines of the treatment process. They might be interested to know the duration of each event in Jane’s journey and how those durations conform to the guidelines of the services.

They might need to detect possible delays in Jane's journey and to know the cause of the delay. They might also need to identify ways to improve the outcome of Jane's treatment.

Those might be easily done for one patient journey, but a health service manager needs to analyse the journeys of all patients within a hospital. Those patients might have various sequences of events in their treatment, depending on many conditions, such as their symptoms, age group, and other characteristics. The hospital might also apply various treatment approaches for different patients that might later result in a different sequence of events or different durations from the same sequence of events. The health service managers need some approaches to analyse those variations. Process mining is one of the promising approaches to analyse durations, sequence of events, variants, and other related analysis.

1.2.3 Process mining perspective

Jane's journey can be represented as a sequence of discrete events: *referral, outpatient consultation, investigation, test result consultation, diagnosis, admission, surgery, after-surgery consultation,* and *discharge*. The clinicians who undertook these events included the gynaecologist, the oncologist, the pathologist, and the surgeon. In each of those activities, the clinicians entered the details of the events in the EHR system. The EHR system added some extra information about the events, for example, the exact timestamp when the event was performed.

A typical EHR system records clinical information in a complex database structure. That includes the details of the event in a specific data format. The event names are stored along with the clinician's login information, and the timestamp when the event happened. Those recorded details are the minimum components of an event log, which can later be analysed to understand the recorded events that have happened within the hospital. The structure of the EHR system is generally more complex than what is needed for process mining. Thus, the first challenge is to extract and transform the details of the events in the suitable format for process mining.

A process mining approach is used in this study to analyse such data. In Jane's example, the treatment process is described in a series of events as actions to achieve the outcome of being cancer free. A systematic approach can be used to select and find the information needed in analysing clinical pathway using process mining approaches. The recorded event name, clinician name, and timestamp of each event can be used to discover the sequence of events that happened to any particular patient. Jane’s journey, for example, can be represented in the following sequence: *GP referral* → *outpatient consultation* → *investigation* → *pathology test* → *MDT review* → *diagnosis* → *test results consultation* → *admission* → *surgery* → *after-surgery consultation* → *discharge.* The sequence of events of many patients might create splitting pathways depending on what is found and what actions are taken.

From the sequence of the events of many patients, a process mining approach can be used to analyse the patient pathways within a hospital. The sequence of events can be compared to the guidelines related to cancer treatment. One example of the guidelines in the NHS is cancer waiting time target, for example, the 62-days wait pathway. This guideline requires the pathways from referral to treatment to be completed in no more than 62 days (two months) [12].

This representation omits things, of course, and this might be worth noting. For example, when Jane and the oncologist discussed and agreed the course of treatment, there might have been a long discussion on how the diagnosis has upset the patient, what options had been given by the clinician, and other considerations. This discussion is not represented in the sequence of events. Some details are possibly recorded in in free text and can be built in to increase the completeness of the analysis in process mining. In this example, process mining only takes account of the sequence itself and not any free text which might, for example contain the reason why the sequence occurred. This would remain as a limitation of this study, as the analysis is carried out based only on the recorded data.

1.2.4 Information system perspective

The EHR system can be seen as a type of information systems where the main users are the clinicians and other healthcare professionals including surgeons and pharmacists. They interact with the information system through the UI. As users work with the system, they input data, do some actions, and get outputs through the UI designed in the system. The UI through which the users interact with the information system plays an important part in the success of the information system. A good UI design can play a key role in the success of the systems, including EHR systems [13].

This can be related to Jane's journey as the example. When Jane came to the hospital, her journey was recorded as data in the EHR system. Some details were typed in manually by clinicians in the hospital or recorded automatically by the EHR system. For example, the admission staff who admitted Jane for surgery would enter Jane's medical information, including the date of the GP referral. The oncologist in Jane's oncology consultation entered more details in Jane's medical information, including the scheduled date for the pathology test. When the oncologist submitted the information, the EHR system would automatically record the time when the oncology consultation took place. The pathologist might have added medical information before/after the pathology test, while the pathology system automatically recorded pathology test results and the time when the pathology test was conducted.

The users of the EHR system in Jane's example were the admission staff, the oncologist, the pathologist, and the surgeon. Those users worked to input data, do some actions, and get outputs through the UI of the EHR system. For example, the admission staff who admitted Jane checked if Jane had already been registered in the system, checked some personal details to make sure that the data are valid, inputted some medical information, and received confirmation from the EHR system that the data had been updated.

Looking back at Jane's journey, all of Jane's treatment was facilitated by the information systems. The admission system was used to record Jane's information on hospital admission and discharge. The oncology system was used by the oncologist to record details of the consultations. The pathology system was used by the pathologist to record pathology test results. The surgery system was used by the surgeon and the surgery team to record information regarding Jane's surgery. In fact, Jane's journey is actually recorded in greater detail than outlined above, for example including when the referral is received, the system makes the appointment, generated the letter, was updated if Jane called to change the appointment, probably resulted in requested per-appointment details, scheduled the time of the appointment, recorded the time that the patient enters the consultation room and the time of the next appointment, etc. Whenever the information system was used, the UI had an important part to play in the success of the EHR system.

For example, when Jane visited the hospital to get admitted, the admission staff needed to check if Jane had already been registered. The UI required the staff to type in a National Health Service (NHS) number, last name, or date of birth to correctly identify Jane. When the data produced a match, the staff were then required to see and verify Jane's personal information. The UI needs to be designed in such a way that personal information can be presented in a suitable layout to make it easy for the staff to see all important information on the screen. The staff might ask Jane some questions to see if any information requires to be updated, for example, Jane's home address. The staff then need to enter the date of the GP referral, along with some other details such as GP name and/or address. The UI of the EHR system needs to be designed such that the date of GP referral can be inputted in a standard format to avoid any error. The regular GP name or address might be presented as a list to improve efficiency.

1.2.5 System change

Over time, the EHR system might change for several reasons. Some users may find that a button is missed out most of the time because it is too small, so the developer team makes it bigger. Other users may find that a menu is placed in the wrong order that does not agree with the order of the care process, so the developer team moves it to the correct order. There might also be a change in the guidelines for patient treatment that requires changes in the records. Another time, the hospital might decide to join an initiative to connect the EHR system to all care records throughout the city, so the developer team makes some changes to adjust the EHR system. In an extreme example, the hospital might decide to change the EHR system to another system completely. All these examples show that the EHR systems have some changes at different levels, for various purposes.

To relate to Jane's example, the EHR system where Jane's journey is recorded might have faced several changes. Those changes might affect the way Jane's journey is recorded in the database of the EHR system. For example, if five years ago, the hospital had a separate pathology system to record all pathology test activities within the hospital. The hospital only started to record pathology test results in the last five years. Whenever a pathology test is done, the EHR system would record the test result and make it available to be updated by the clinicians, as required. If Jane had come five years earlier, her records in the EHR system would not include any activities related to the pathology tests. In this case, the analysis would reveal that Jane's journey was not complete because she did not have a pathology test. This does not mean, of course, that Jane had been diagnosed without a pathology test but illustrates an effect of a system change over time.

This thesis focused on the changes in the UI of EHR systems. The aim is to build a method to examine and analyse different types of UI changes and the effect of those changes on the processes. The process change is analysed using a process mining approach. Process mining is used in this study because it can be useful in analysing and understanding processes through the event log automatically generated in the information system. The current literature in process mining has shown promising results in the analysis and understanding of the processes. Still, there are limited studies discussing the effect of system changes on process changes over time. This study demonstrates the suitability of using process mining for process change analysis within the healthcare domain, to understand the suitability for using EHR given changes over time and how to control for the change over time. The motivation of this thesis is to make contributions to the healthcare process mining community [14], where one of the main challenges is the complexity due to the changes over time.

1.3 Objective, hypothesis, and research questions

The objective of this research is to detect and analyse process changes in EHR systems. In aiming to meet this objective, the key hypothesis is that process mining can be used to analyse process change in the EHR system. The Research Questions are developed by breaking down the key hypothesis, as presented in Figure 1.1.

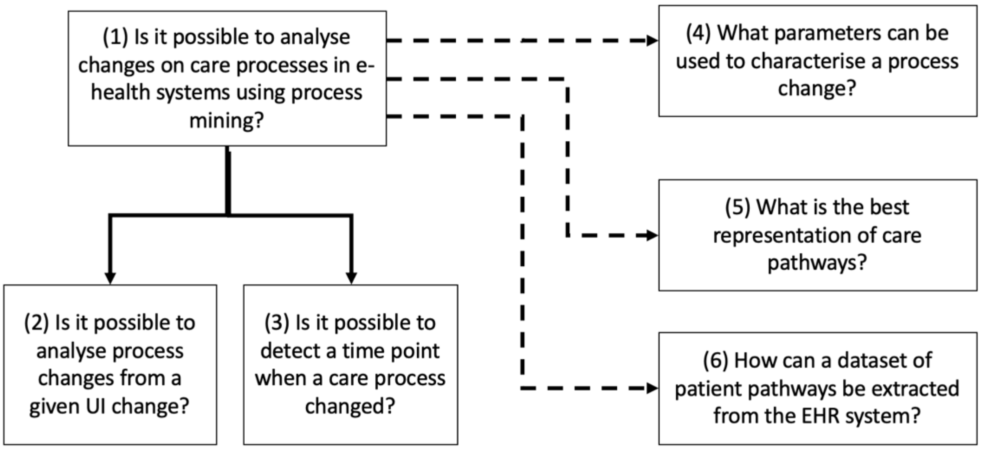

(1) Is it possible to analyse changes on care processes in e-health systems using process mining?
(4) What parameters can be used to characterise a process change?
(5) What is the best representation of care pathways?
(2) Is it possible to analyse process changes from a given UI change?
(3) Is it possible to detect a time point when a care process changed?
(6) How can a dataset of patient pathways be extracted from the EHR system?

Figure 1.1 Research Question development. The left side shows the primary questions with two possible cases connected with solid arrows. The right side shows the measurement questions in either one of two cases (2 or 3 on the left), connected with dashed arrows.

The first research question is “Is it possible to analyse changes in the care processes in an e-health system using process mining?” (RQ-1). This research question can be split into two based on the pre-condition whether a process change point is known or not: “Is it possible to analyse process change from a given UI change?” (RQ-2) or “Is it possible to detect a point in time when a care process changed?” (RQ-3). In a case where UI changes have been documented, it may be possible to know the exact point in time when the change happened. If this is the case, the analysis focuses on how a given UI change affects the care process. Another case is where a care process is undertaken over a long period, the process might have been changed over time. In this case, the analysis focuses on how to characterise this process change over time.

Further, Figure 1.1 shows that this research needs to find parameters that can be used to characterise process changes (RQ-4). Those parameters can be defined based on process characteristics. A process can be characterised from many representations and perspectives, for example, based on the sequence of activities, based on the duration, or based on the outcomes. Each of those representations and perspectives may change over time. To support this question, this research needs to consider how best to represent the care pathways (RQ-5). The fundamental question is how to extract an event log, a dataset of patient pathways, from the EHR system (RQ-6). The database was not developed from a process perspective; hence the initial challenge is to get the correct data required to analyse processes.

1.4 Study approach

The general approach of this study was to explore the dataset of patient pathways as extracted from the EHR system, test the hypothesis through some experiments of process change analysis, and evaluate the results through discussions with clinical experts. The main inputs of this research were the EHR database, the information about a UI change, and advice on clinical treatments. The EHR database was explored to extract the data related to the process change. Using process mining approaches, those data were transformed to represent and characterise the process as needed by clinical experts to better understand the process change. Based on those representations and characteristics, the analysis was completed by discovering process changes over time. The output was presented in many visualisations to support understanding of the process.

The methodology in this study is adapted from the two well-established process mining methods, the L* life-cycle model [3] and the Process Mining Project Methodology (PM^2) [15], and they are discussed in more detail in Chapter 3. The study started with planning and justification to define the scope, research questions, and the data of this study. It was followed by Extraction, Transformation, and Loading (ETL) of the identified data required for this study. The next stage is the main part of this study, where the process mining and process change analysis was done. The process change analysis [16] was conducted by partitioning the event log over time and by process mining each of them to discover the process models over time. The last stage is evaluation, which was conducted through both a statistical evaluation and a clinical evaluation. The statistical evaluation was done using hypothesis testing, to test if there is a statistically significant change before and after a change point in time. The clinical evaluation was done through focused group discussions with clinical experts to unravel the real changes happening within the EHR system.

Based on the study planning and justification, the datasets used in this study were chosen from two data sources to understand the variability of the EHR systems. The datasets came from two data sources: one from a hospital in the United States of America (USA) and one from a hospital in the United Kingdom (UK). The USA represents a country with a non-universal healthcare insurance system where healthcare services are largely provided by private providers [17]. The UK represents a country with a universal government-funded health system, also known as *single-payer healthcare* [18]. By analysing datasets from the USA and the UK, this study may be extended to an international comparison of healthcare services. The data provenance of those two data sources are described in the following section.

1.4.1 Data sources

There are two data sources for this study. These are: the Beth Israel Deaconess Medical Center (BIDMC), USA, and the St James's University Hospital (SJUH), UK. Those two data sources come from hospitals located in two cities of comparable size. Without trying to compare those two hospitals with one another, the following sections present the overview of those two data sources.

1) The Beth Israel Deaconess Medical Center in Boston, USA

The first data source was the EHR of the BIDMC hospital in Boston, USA. This hospital was formed in 1996 by a merger of Beth Israel Hospital (founded in 1916) and New England Deaconess Hospital (founded in 1896). BIDMC is a private hospital for Harvard Medical School and is located in Boston [19].

Boston is the capital and the most populous city of Massachusetts in the US. The city covers 89.63 square miles (232.14 km^2) with an estimated population of 694,583 in 2018. In addition, there are hundred thousands of people who travel to Boston for work, education, healthcare, and special events. The population rises to 1.2 million during working hours and 2 million during special events. Boston is a home to an affluent population and is a wealthy city with one of the highest costs of living in the USA[20].

In March 2019, the BIDMC joined Beth Israel Lahey Health, a new healthcare system in the US, along with 11 other hospitals. Since the merger, the hospital has consisted of two campuses, the East (former Beth Israel) and the West (former Deaconess). The East Campus houses most of the primary care, outpatient, clinical and administrative functions. The West Campus retains the department of human resources, the emergency department, inpatient care, and many specialists. The medical center has more than 6,000 full-time employees. The BIDMC provides 673 licensed beds, which includes 493 beds in medical/surgical department, 77 beds in critical care and 62 beds in obstetrics/ gynaecology (OB/GYN) [19].

2) The St James's University Hospital in Leeds, UK

The second data source for this study was the EHR of the SJUH in Leeds, UK [21]. This public hospital was originally the Leeds Moral and Industrial Training School built in 1848 and was named as St James's University Hospital in 1970, which is located in Leeds, UK.

Leeds is a city in West Yorkshire, England, UK. Leeds has one of the most diverse economies and the fastest rate of growth in the UK. The city covers 213 square miles (551.7 km^2) with an estimated population of 789,194 in 2018. Leeds economy is one of the most diverse of all the UK main employment centres. Leeds is one of the largest business centres in the UK with around a quarter million people are employed in the Leeds City Region in the financial and professional sector [22].

Along with five other hospitals, SJUH is managed by the Leeds Teaching Hospitals NHS Trust (LTHT). The LTHT is the largest provider of specialised services in England [23]. The SJUH provides 997 beds, including the new oncology building, the Bexley Wing. The Leeds Cancer Centre is one of Europe's largest cancer centres with 350 beds and 1,600 staff [24].

1.4.2 Overview of the datasets

The EHR data from those two data sources were made accessible to the researcher. Initially, access was gained to the open accessible dataset from the BIDMC, USA. This is called the Medical Information Mart for Intensive Care III (MIMIC-III). The second dataset was the de-identified subset of the Patient Pathway Manager (PPM) data from the LTHT. This is called the PPM Chemotherapy dataset. Subsequently, the third dataset was accessible from the full access of PPM dataset. This is called the PPM Cancer dataset. The second and third datasets came from the same data source, but were considered as two different datasets due to their data access, anonymisation process, and the scope of the datasets.A summary of the datasets in this study is presented in Table 1.1.

Table 1.1 Data summary

Dataset	Hospital	EHR System	Experiment
1) MIMIC-III (n=46,520)	BIDMC, Boston USA	- Carevue - Metavision	1. Colorectal cancer patients (n=1,600) 2. All patients (n=46,520)
2) PPM Chemotherapy (n=31,511)	LTHT, Leeds UK	Patient Pathway Manager (PPM)	Breast cancer patients receiving EC-90 adjuvant chemotherapy between 2003 – 2012 (n=738)
3) PPM Cancer (n>2.5M)			1. Breast cancer patients receiving EC-90 adjuvant chemotherapy between 2014-2018 (n=733) 2. Referral to diagnosis of Leeds patients diagnosed with endometrial cancer (n=1,664) 3. Referral to first treatment of Leeds patients diagnosed with endometrial cancer (n= 949) 4. Cancer patients

** n = number of patients*

Those three datasets were carefully chosen to represent variability of EHR data. The MIMIC-III dataset includes patients in critical care units only. MIMIC-III was chosen because this is publicly available and supporting reproducibility of the research using the same dataset. The PPM Chemotherapy dataset is an extract of patient data consisting data of patients receiving chemotherapy in the LTHT. This dataset represents a specific cohort of patients following a general pathway of chemotherapy treatment. The PPM Cancer dataset is the full dataset of cancer patients in the LTHT. This dataset represents a larger cohort of patients with many different pathways followed. Analysis of the PPM Cancer dataset was also supported with adirect access to representatives of clinicians and PPM developers. For simplicity and consistency reasons, the dataset names to be used in this research are the MIMIC-III dataset, the PPM Chemotherapy dataset, and the PPM Cancer dataset. Case studies of these datasets are described in Section 3.4 and are explored in more detail in Chapters 4–6.

1.5 Thesis structure

This thesis is structured based on the experiments conducted on the three datasets as three separate case studies. The case studies are presented as separate chapters but experiments are numbered chronologically based on the order on which they were conducted. The structure of this thesis and description of the chapter headings are as follow.

1) Introduction

This chapter introduces the importance of improving healthcare processes through data. An illustrative example of a patient journey is presented to set a problem definition which can be seen from different perspectives i.e. the health service manager, process mining, and information system perspectives. The UI change is then introduced to highlight the main focus of this study. This chapter is concluded with the establishment of research objectives, research questions, approaches, and the thesis structure.

2) Background

This chapter summarises the literature in healthcare and technical areas as the background for this thesis. The healthcare background includes a review of healthcare systems, electronic health record research, coding standards, process guidelines, and cancer. The technical background includes workflow technology and process modelling, process modelling notation, process mining, process mining in healthcare, process mining to analyse process changes, and statistical approaches.

3) *Methodology*

This chapter describes the general methodology and the three datasets used in the study. The four main steps in the general methodology following the PM^2 are: (1) Plan and Justify; (2) Extract, Transform, Load; (3) Mine and Analyse, and (4) Evaluate. The three datasets used in this study are the MIMIC-III data, the PPM Chemotherapy data, and the PPM Cancer data. Each dataset will be treated as a separate case study and will be discussed in a dedicated chapter. Chapter 3 describes the general methodology, while Chapters 4–6 are organised around the three datasets.

4) *Case study 1: Experiments using the MIMIC-III dataset*

The first part of this chapter explores the MIMIC-III dataset in terms of the data characterisation, data provenance, scope, representativeness, data quality, data variety, and the limitations.

The second part of this chapter describes how process mining has been used to analyse the data from the MIMIC-III dataset, following the steps in the general methodology. It starts with the description of cancer patient treatments in the BIDMC data source; followed by introducing a new stage called the Database Reconstruction; ETL; process discovery and conformance checking; and comparing processes in CV and MV systems.

5) *Case study 2: Experiments using the PPM Chemotherapy dataset*

The first part of this chapter explores the PPM Chemotherapy dataset in terms of the data characterisation, data provenance, scope, representativeness, data quality, data variety, and the limitations.

The second part of this chapter describes how process mining has been used to analyse the PPM Chemotherapy database, following the steps in the general methodology. It starts with the description of PPM Chemotherapy data; followed by ETL; process discovery and conformance checking; and the process analytics. The process analytics include process mining to reproduce data analysis in chemotherapy process, trace clustering for similarity analysis, and the detection of system changes and their effects on treatment processes in the PPM EHR system.

6) *Case study 3: Experiments using the PPM Cancer dataset*

This chapter explores the PPM Cancer dataset in terms of the data characterisation, data provenance, scope, representativeness, data quality, data variety, and limitations.

This chapter describes how process mining has been used to analyse the PPM Cancer data, following the steps in the general methodology. It starts with the description of PPM Cancer data from the PPM database; followed by ETL; process discovery and conformance checking; and the process analytics. The process analytics include the windowing methods, process comparison and the detection of system changes and their effects on treatment processes in the PPM EHR system.

7) Discussion

This chapter explores the challenges on healthcare process mining, the data analytics to analyse processes, process change analysis, the effect of system change in healthcare processes, and the contributions of this thesis. The challenges on healthcare process mining include data access and ethics approval, data quality, data understanding, and the data and process visualisation. Contributions of this thesis include using the process mining approach on three datasets, exploring the dimensions of process change analysis, and time window selection to analyse the process.

8) Summary

This chapter concludes the whole study with the findings in terms of the method developed and the case studies explored in the study. This chapter also identifies future work to be done based on the findings and lessons learned in the study.

1.6 Summary

This first chapter has introduced the research undertaken during this PhD, by providing an overview and an illustrative example in order to understand the problem domain. The example has been explored from the perspectives of health service, process mining, and information systems. UI change has been introduced as one main challenge to be explored in this study. The aim, research objectives, research questions, approaches, and the thesis structure have also been presented. The next chapter will review the literature which builds the background for this study.

There is an opportunity to improve healthcare through the application of process mining to EHR data, but there are some challenges as described. The impact of UI changes on the data is one of these challenges. The case study method will be adopted based on the pragmatic availability of data to test the hypothesis that process mining can be used to understand process change over time.

Chapter 2
Background

Chapter 1 introduced this thesis as a study that needed both healthcare and technical expertise. Here in Chapter 2, the literature within the healthcare and technical background is reviewed to develop the context of the study. This chapter includes a jointly-authored publication presented in the IEEE International Conference on Information Communication and Management (ICICM) entitled "Process mining in oncology: a literature review" [25]. This literature review paper is summarised in Section 2.2.4.2.

2.1 Healthcare background

The relevant background to healthcare is presented in this section. Healthcare is a complex domain that includes the prevention, diagnosis, and treatment of disease, injury, illness, and other conditions in people [26]. Healthcare is delivered by health practitioners or providers i.e. primary care, secondary care, tertiary care, community care, and public health. Healthcare professionals include doctors, nurses, therapists, pharmacists. The following subsections have been structured to describe the general introduction to healthcare systems, how the healthcare data has been used in the EHR research, some coding standards and process guidelines related to this study, and a description of cancer as a specific domain analysed in this study.

2.1.1 Healthcare systems

A healthcare system is the organisation of people, institutions, and resources that deliver healthcare services to meet the health needs of the populations. According to the World Health Organization (WHO), a healthcare system consists of all organisations, people and actions to promote, restore, or maintain health. The goals of healthcare systems are good health for the populations, responsiveness to the expectations, and fair funding [27].

Healthcare systems vary across countries [27]. They can be classified based on the funding body supporting the healthcare system. Some countries with universal government-funded health systems are Australia, Canada, and UK. Some countries with universal public insurance systems are China, Japan, and United Arab Emirates. Some countries with universal public-private insurance system are Austria, Chile, and Germany. Some countries with universal private health insurance are Netherlands and Switzerland. Some countries with non-universal insurance systems are Egypt, Indonesia, and the USA. This thesis analyses healthcare data from the USA and UK as two countries with different type of healthcare systems.

The healthcare system in the USA is a non-universal insurance system, where healthcare is operated under a mixed market healthcare system. It is mostly funded by private health insurance and a small number are funded by public health coverage such as Medicare, Medicaid, and the Veteran Health Administration [28]. Some states in the USA are moving towards universal healthcare coverage, including Minnesota and Massachusetts. In this study, the USA healthcare system was explored to analyse the MIMIC-III data from the Beth Israel Deaconess Medical Center (BIDMC) hospital. No additional information was required in order to perform the analysis of the dataset.

The healthcare system in the UK is a single-payer system, where healthcare is controlled by the government, funded by leveraging taxes and redistributing across the population, and made available to all citizens [29]. In this thesis, the UK healthcare system was explored by using data from the Leeds Teaching Hospitals NHS Trust (LTHT) as a secondary/tertiary care provider. Relevant details of the healthcare system were discussed with clinical experts during the experiments using the datasets.

2.1.2 Electronic Health Record research

Information is one of the main components of a healthcare system. The Electronic Health Record (EHR) is the computerised format of healthcare information, which stores the health-related data of all patients treated within a healthcare institution, which may include records of patients’ referral, symptoms, past medical history, physical examination, diagnosis, tests, procedures, treatment, medication, and discharge [30]. An EHR contains the clinical histories of patients, which are stored to support the clinical care delivery for patients, improve the performance of the clinical practice, and facilitate clinical research. In any implementation, an EHR can be fully integrated or partially integrated in the healthcare institution.

The EHRs in general have been used for many research projects, including epidemiology [31], observational research, safety surveillance and regulatory uses, and prospective clinical research [32]. The benefits of using EHRs in clinical studies include, among others, integrating large amounts of medical information, enabling a longitudinal study of diseases, and enabling linkage to other datasets. The actual benefits of a study using EHRs compared to other studies are that (1) savings in cost, time, and labour are possible because additional data collection is not needed, and (2) it supports decision making based on the recorded data in the EHR, which is known as evidence-based decision making. The challenges to using EHRs in research are related to data quality, complete data capture, and heterogeneity between systems. From a pathway point of view, the challenges are that the EHR data recorded for clinical purpose might contain much more details that are not needed in the analysis, not recorded complete sequence of patient pathway, recorded in different levels of detail than needed, or might spread in different systems with different recording formats. The first challenge to work with the EHR data is therefore how to find the suitable data needed for the analysis.

One interesting type of observational research in the EHR systems is clinical pathway analysis. A clinical pathway is a sequence of events in the patient care with a specific clinical problem [8]. The purpose of clinical pathway analysis varies from analysing patterns and deviations of the clinical pathway to analysing outcomes. Analysing patterns of clinical pathways is important in understanding medical behaviours and the order of activities in patient treatment [33]. Analysing outcomes are important to recommend procedures and treatments suitable for patients to gain positive health outcomes. By observing the recorded data in the EHR systems, pathways analysis can be supported by the recorded data as the evidence of any findings.

Process mining is a promising approach in order to analyse the patterns of clinical pathways based on the data recorded in the EHR system. It is particularly interesting as a secondary use of routinely collected data in the healthcare settings to understand and improve health services [34]. The main input for process mining is an event log, which records the events in a pathway as they are undertaken by clinicians in a clinical setting. The models discovered by analysing the event log are representatives of the pathways being analysed. More details about process mining and how it can be used for clinical pathway analysis is presented in Section 2.2.3.

2.1.3 Coding standards: the International Classification of Diseases

There are a range of coding standards used worldwide, but the International Classification of Diseases (ICD) is the most widely used for classifying diseases. It is mandated by the World Health Organisation (WHO) for the statistical analysis of diseases and mortality. ICD is a standard for the diagnostic classification that can be used for ensuring coding consistency in clinical and research purposes. [35]

The first version of this system was introduced in 1893. After several changes, the ICD-9 (1978) was introduced to revise the standards and to reflect advances in health and medical science over time. The ICD-10 (1994) allowed for significantly more codes and permitted the tracking of many new diagnoses and procedures. The ICD-10 is an extension of the ICD-9 with additional codes. It is possible to map codes from ICD-9 to ICD-10. The latest version, ICD-11, was released in June 2018, but it is not currently used in practice. [36]

In this study, the MIMIC-III data used the ICD-9 codes while the PPM data used the ICD-10 codes. Understanding of both ICD-9 and ICD-10 codes is important in order to support analysis of specific cohorts of patients.

2.1.4 Process guidelines

Several guidelines are used in healthcare to standardise processes within the EHR system. Some of these are followed internationally, nationally, and locally. The process guidelines followed in this study based on the UK healthcare system are the National Institute for Health and Care Excellence (NICE) guidelines as national guidelines in the UK and the Leeds Health Pathways as local guidelines in the LTHT. Process guidelines in the US healthcare system are not discussed because there was no direct access to the BIDMC hospital to confirm which ones were followed.

The NICE pathway guidelines are evidence-based recommendations for health and care in England. The NICE is an executive public body in the United Kingdom, a part of the Department of Health that provides national guidance and pathways to conduct health and social care in England and Wales. NICE assesses the recommended treatment of diseases medically and economically [37]. For example, there is specific guidance based on the condition of breast cancer, more specifically in advanced (stage 4) breast cancer [38]. The pathway is presented as an interactive flowchart, which includes information and support, imaging assessment, pathological assessment, and the management of a person with suspected advanced breast cancer. The NICE guidance is a high level standard that needs to be broken down into a detailed level suitable to specific practices in hospitals.

The Leeds Health Pathways (LHP) [39] are a set of local guidelines adopted by the LTHT and other organisations including the Yorkshire Cancer Network. They were adopted from the guidelines from various professional bodies including NICE. The LHP is provided as part of the clinical guidelines in the LTHT for patient treatment in various diseases. In gynaecology, for example, it provides guidelines for abnormal bleeding (post-menopausal/pre-menopausal), management of premenstrual syndrome, and lumps and bumps, among others. The guideline includes background information or scope of pathway, information resources for patients and carers, development and updates to the pathway, and referral forms. More details referring to the LHP about specific types of cancer are presented in Sections 2.1.6.

Understanding the guidelines and ontology of the healthcare domain is a part of this study, as a reference in discovery and conformance checking steps. Generally, ICD codes are used to code diagnoses and procedures, while the NICE pathways and the LHP are useful as guidance of patient treatment. It is important to check if the discovered process models are conforming to the guidelines used in the healthcare processes to assess the quality of the guidelines and possibly find best practices.

2.1.5 Cancer

Cancer is a leading cause of death worldwide. There were an estimated 9.6 million deaths caused by cancer in 2018 [40]. The five most common causes of cancer death are breast, colorectal, liver, lungs and stomach cancers. In the UK, cancer is the fourth leading cause of death [12]. There are around 164,000 cancer deaths every year (2014–2016), with the most common being lung, bowel, breast, and prostate cancers [41].

Cancer is a group of diseases involving abnormal cell growth. Cancer has the potential to invade or spread to other parts of the body [42]. Other terms used are *malignant tumours* and *neoplasms*. Cancer is a genetic disease caused by changes to the genes that control how cells grow and divide. A cancer that spreads from the original place to another place in the body is called *metastatic* cancer [43]. Most cancers are recognised through the presenting symptoms or through screening. The definitive diagnoses would require the examination of a tissue sample by a pathologist. There are more than 100 types of cancer, which are named from the organs or tissues where the cancers form.

The general guidance for cancer contained in the NICE guidelines is NG12 for suspected cancer: recognition and referral [44]. This guideline outlines appropriate investigations in primary care on children, young people, and adults with symptoms that could be caused by cancer. The recommendations are organised by the site of the suspected cancer, the symptom, and the findings of primary care investigations. The guideline is provided for healthcare professionals, people involved in clinical governance in both primary and secondary care, and also for people with suspected cancer and their families and/or carers.

2.1.6 Cancer types

The following sections discuss the three types of cancer analysed in this study, which are colorectal, breast, and endometrial cancers. The selection was based on discussions with clinical experts to get representative cohorts of cancer patients in those three datasets.

2.1.6.1 Colorectal cancer

Colorectal cancer is cancer that develops from the colon or rectum (part of the large intestine). It is also known as bowel cancer, colon cancer, or rectal cancer. The risk of getting colorectal cancers is increased by old age and lifestyle factors. Less than 5% of cases are due to genetic disorders. The five year survival rate in the United States is around 65%. Globally, colorectal cancer is the third most common type of cancer, making up about 10% of all cases [45].

Diagnoses may be obtained through a physical exam and history, faecal occult blood test (FOBT), x-rays, sigmoidoscopy or colonoscopy, followed by medical imaging [46, 47]. Surgery is the most common treatment for all stages, especially for cancers that are confined within the wall of the colon. The other types of treatment such as chemotherapy and immunotherapy are undertaken in specific conditions. In people with incurable colorectal cancer, palliative care is recommended. Palliative care can consist of procedures to relieve symptoms or complications from the cancer but do not attempt to cure the underlying cancer.

2.1.6.2 Breast cancer

Breast cancer is the development of cancer from breast tissue. Globally, breast cancer affects about 12% of women and is the most common cancer in women. The risk of breast cancer is increased by many factors, including a family history of breast cancer. It is sometimes caused by BReast CAncer (BRCA) gene mutations [48].

There are 34 guidance points, 4 pathways, 2 quality standards, and 8 advice points contained in the NICE guideline for breast cancer. Most of them are related to specific treatment or diagnostics guidance. The general guidance topics are: Advanced breast cancer diagnosis and treatment (CG81) [49]; Familial breast cancer (CG164) [50]; and Suspected cancer recognition and referral (NG12) [44]. The LTHT guidelines for the treatment pathway of breast cancer is started with a referral from a GP or other source [51]. Breast cancer is diagnosed by physical exam, mammogram, ultrasound, MRI, blood chemistry studies, and biopsy of the affected area of the breast. Surgery is the main treatment for breast cancer, which may be followed by chemotherapy or radiation therapy, or both. There might also be adjuvant therapy and neoadjuvant therapy applied. Adjuvant therapy is drug used after and in addition to surgery, while neoadjuvant therapy is chemotherapy or other types of therapy before surgery.

2.1.6.3 Endometrial cancer

Endometrial cancer is a uterine cancer that begins in the inner uterine lining (endometrium) because of an abnormal growth of cells that then can spread to other parts of the body. This is the third most common cause of death in cancers (newly occurred in 320,000 women in 2012) which only affect women, after ovarian and cervical cancer. This mostly occurs in women between the ages of 60 and 70. The most frequent type of endometrial cancer (80%) is endometrioid carcinoma [52].

There is only one NICE guidance point specific for endometrial cancer, that is the laparoscopic hysterectomy for endometrial cancer, published in September 2010 [53]. This is a surgical procedure to remove the uterus. The more general guideline related to endometrial cancer is the NICE guideline NG12 for suspected cancer: recognition and referral [44]. The LTHT guidelines for the treatment pathway of endometrial cancer is started with a referral from a GP to post-menopausal bleed clinic [54]. It is triaged by the benign Gynaecology team. Some tests are done including ultrasound and hysteroscopy. If cancer is confirmed, the most common treatment is surgery, with consideration for radiotherapy or chemotherapy for special cases.

2.1.7 Cancer waiting times

One important performance indicator for cancer treatment in the UK is the cancer waiting time. The NHS has set maximum waiting time standards for cancer treatment. The achievement of the national cancer waiting times is considered to be an indicator of the quality of cancer diagnosis, treatment, and care. There are some standards for waiting times for cancer [12], including:

1. *14-day (two weeks) wait.* This requires patients who had an urgent GP referral for suspected cancer to be first seen by a specialist within two weeks. This is targeted to support early diagnosis in order to spot cancer early and improve survival.
2. *31-day wait.* This requires patients to receive their first cancer treatment within 31 days of a decision to treat. It is also required for a maximum 31-day wait for subsequent treatment where the treatment is surgery, or is a course of radiotherapy, or is an anti-cancer drug regimen.
3. *62-day wait.* This requires patients to wait for no more than two months (62 days) between the date of an urgent GP referral for suspected cancer or an NHS cancer screening service and the first definitive treatment for cancer. There are also requirements for a maximum 62-day wait for the first definitive treatment following a priority upgrade of the patient for all cancers.

In the UK, cancer waiting times data are collected from NHS providers in monthly and quarterly reports. These reports are used to monitor cancer waiting times targets and plan service improvements. Summary of overall performance against all cancer waiting time standards in 2018-2019 ranged from 79.1% to 99.3%. The lowest performance (79.1%) was for the 62-day wait for first treatment following an urgent GP referral for all cancers. The highest performance (99.3%) was for the 31-day wait for second or subsequent treatment on anti-cancer drug treatments [55].

Understanding cancer waiting times standards is important in this study as a starting point to understand the important quality indicator of cancer treatment. Pathway analysis with a process mining approach is useful to examine factors contributing to the achievement of these standards.

2.2 Technical background

The literature reviewed in this technical background section starts with an overview of workflow technology and process modelling. Workflow technology is the root of process mining, both aim to analyse the sequence of activities in a process and create a process model based on the sequence. It is then followed with a focus on process mining, process mining in healthcare, and process mining to analyse process change analysis. This section is closed with a presentation of the statistical approaches used in this study.

2.2.1 Workflow technology and process modelling

Workflow technology was formalised by the Workflow Management Coalition (WfMC) through the Workflow Reference Model [56]. The WfMC was founded in 1993 and is a global organisation of adopters, developers, consultants, analysts, as well as university and research groups working in workflow and BPM. The formalisation of the Workflow Reference Model was followed by the Workflow Patterns initiative, started in 1999 [57], as a conceptual basis for workflow technology and process modelling.

A workflow management system is a system that defines and manages processes through the software whose order of execution is driven by a computer representation of the workflow process logic [60, 62]. Historically, models of the workflow have spanned from the very informal to the very formal. Other pieces of literature use the term “business process” or “process” to refer to “workflow”.

There are many approaches for representing processes, such as state transition diagrams [59], Unified Modelling Language (UML) [69, 70], Business Process Modelling Notations (BPMN) [62] and Petri nets [63]. All of these process model types are useful for different purposes, for example, to view the process from various angles, to structure discussions among stakeholders, to analyse performance, or to “play out” different scenarios and provide feedback. Each of these four process modelling notations (state transition diagram, UML, BPMN, and Petri nets) are described in more detail in Section 2.2.2.

2.2.2 Process modelling notations

In process modelling, many notations can be used, including transition systems, UML activity diagrams, BPMNs, and Petri nets. Each of these notations is described in this section along with an example for a typical pathway to illustrate the similarity and differences.

1) **Transition system diagram**

The most basic process modelling notation used is a transition system or state transition diagram. It consists of states and transitions, which may be marked with labels chosen from a set [64]. The formal definition of a transition system is a set of states, activities, and transitions (*S, A, T*) [4]. A state is an identifier with a unique label. A transition connects two states and is labelled with the name of an activity. A transition from state *p* to state *q* is written as *p* → *q*. The states are represented by black circles, the actions by directed arcs, and the transitions connected two nodes by an arc. Multiple arcs can have the same label.

Figure 2.1 shows an example of a transition system modelling six defined activities in a hospital administration process of a cohort of patients. The formal definition is as follows: *S = {S1, S2, S3, S4, S5, S6, S7, S8}*, S^{start} *= {S1}*, S^{end} *= {S8}, A = {ED reg, admission, ED out, discharge, death, death|discharge}* and *T = {(S1, ED reg, S2), (S2, ED out, S3), (S2, ED out, S4), ..., (S7, discharge, S8)}*.

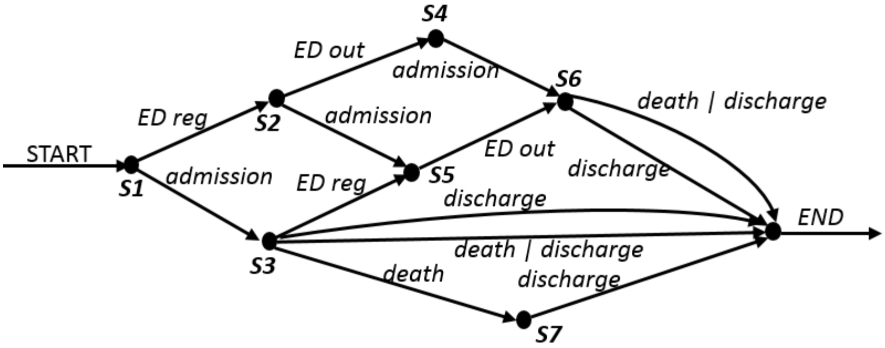


Figure 2.1 A transition system example. Transitions represent activities in the process flowing from one state to the other. The modelled process flows from START to END.

Any process model with executable semantics can be mapped into a transition system. Transition systems are simple but have limitations to express concurrency [65]. Given the concurrent nature of business processes, there are a variety of options for more expressive models.

2) Unified Modelling Language Activity Diagrams (UML ADs)

UML AD is one of the UML diagram types, which consist of structural diagrams and behaviour diagrams. The UML structural diagrams represent the static view of a system model, which include a class diagram, component diagram, and deployment diagram. The UML behaviour diagrams represent the dynamic view of a system model, which include a sequence diagram, activity diagram, and state machine diagram [61]. All of these diagrams were created to standardise the design of a system.

A UML AD is constructed from a number of shapes connected with arrows. The actions are represented by rounded rectangles, decisions represented by diamonds, the start (split) or end (join) of concurrent activities represented by bars, the start (initial state) of the workflow represented by a black circle, and the end represented by an encircled black circle. In process modelling, actions are also known as activities. There is also a possibility to represent multiple actors involved in a process through the swim lanes. A swim lane is a visual representation of a route through the activity diagram which represents the activities performed by a particular set of actor(s).

Figure 2.2 shows a UML activity diagram of the same hospital administration process. This diagram consists of six actions (*admission, ED reg, ED out, death | discharge, death, discharge*) with one pair of concurrent activities, and three decisions.

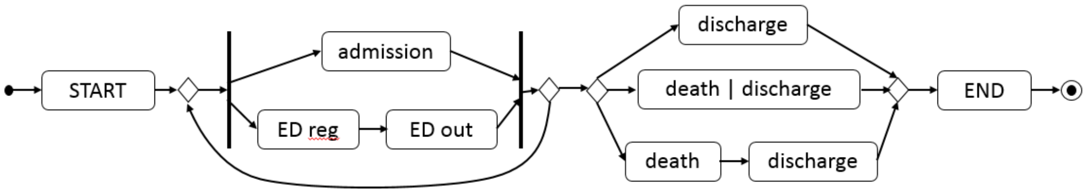


Figure 2.2 A UML activity diagram. Rounded rectangles represent the activities, diamonds represent the decisions. The modelled process flows from a black circle as the START to a encircled black circle as the END.

One advantage is that this diagram follows the UML modelling language that is intended to provide a standard way to visualise the design of a system. It can be seen as a combination of a structured flowchart with a traditional data flow diagram. If a typical flowchart cannot express concurrency, the join and split symbols in the UML activity diagrams can resolve this. Therefore, it is easily understood by both analysts and stakeholders. This diagram can also be mapped into a state transition diagram or a flowchart [66].

3) Business Process Modelling Notation (BPMN)

BPMN was developed by the Business Process Management Initiative (BPMI) and is maintained by the Object Management Group (OMG). BPMI was established in 2020 and is a non-profit organisation that promote the standardisation of common business processes. OMG was founded in 1989 and is a consortium that develops a heterogeneous distributed object standard. BPMI and UML were merged in 2005. BPMN was adopted as a standard by OMG in 2006. BPMN is a standard for capturing business processes in system development. [62].

An example of a process model in the BPMN notation is shown in Figure 2.3. This process model represents a hospital administration process of a cohort of patients. The obvious differences compared to the UML AD are in the notations that represent the start and end actions, decisions, split and join (and/ or). The use of gateways in BPMN makes it simpler than the use of diamonds and bars in the UML AD.

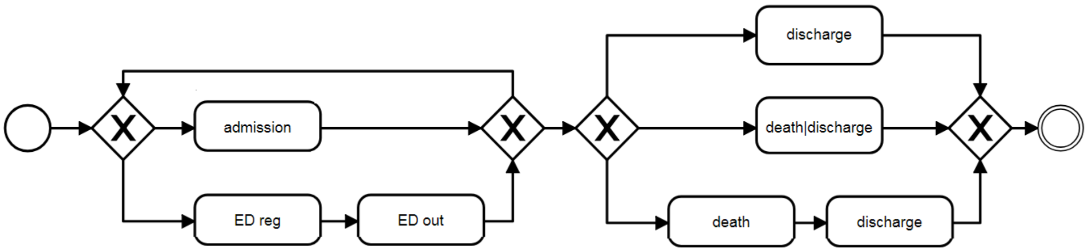


Figure 2.3 A BPMN process model. Rounded rectangles represent the activities, gateways represent the decisions.

Both BPMN and UML are managed by OMG, which make both diagrams very similar to each other. Both BPMN and UML AD were developed to be equally easy to understand by the analysts and the stakeholders [67]. BPMN emphasises the possibility to model different events and exceptions for routing a process. This makes BPMN is suitable to model clinical pathways in healthcare domain [68, 69].

4) Petri Nets

A Petri Net is a process modelling technique invented by Carl Adam Petri [63]. Since then, Petri Nets have been used to model and analyse many kinds of processes. Petri Nets are widely used for workflow modelling [75, 76] because of the clear and precise formal semantics, intuitive graphical nature, basic properties, and the availability of many analysis techniques.

An example of a Petri Net showing the hospital administration process of a cohort of patients is presented in Figure 2.4. It consists of places and transitions, connected by arcs from a place to a transition or vice versa.

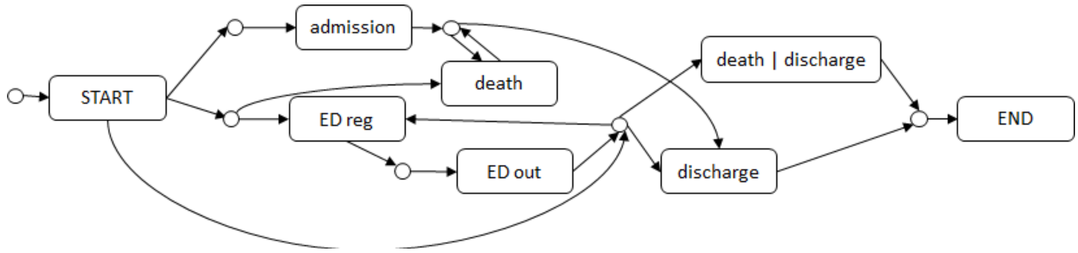


Figure 2.4 A Petri Net. Empty nodes are places and rounded rectangles are transitions that represent the transitions of activities in the event log.

As described above, a large number of process modelling notations can be used as care pathway modelling techniques in healthcare research. Process models can be used for many purposes, such as: to structure discussions with stakeholders from several backgrounds; to support documentation; to verify and find errors in a process; to analyse performance; or to configure a system [4]. In this study, process models are used in the process discovery task of process mining based on the data recorded in the event log.

2.2.3 Process mining

The development of workflow technology was continued by the establishment of the Institute of Electrical and Electronics Engineers (IEEE) Task Force on Process Mining in 2009 [3]. This task force promotes the research, development, education and understanding of process mining. The term “process mining” was coined by van der Aalst to describe a specific type of workflow analysis and has been used widely since 2003. Process mining joins the ideas of process modelling and analysis on one side and data mining and machine learning on the other side.

The idea of process mining is to discover, monitor, and improve real processes by extracting knowledge from event logs readily available in the information systems [72]. The processes being analysed are the real processes as they were recorded in the information systems, not the assumed processes. The goals of process mining are to detect previously unknown process structures, to analyse the occurrence of process pathways in the system, or to quantify the conformance of the process to guidelines [73]. Three general steps of process mining are process discovery, conformance checking, and enhancement. Each of those will be explored in this section.

Process mining can be done from several perspectives [74]:

- The control-flow perspective focuses on the ordering of activities.
- The organisational perspective focuses on how resources (e.g. people, systems, roles, and departments) are involved and related.
- The case perspective focuses on the properties of cases. A case can be characterised by its path in the process, by the actors working on it, or by the values of the corresponding data elements.
- The time perspective is concerned with the timing and frequency of events.

The main focus of this research is the control-flow perspective to understand the patterns of activity sequences. However, an additional perspective is needed in this research, which is the time perspective, to analyse the process change over time.

2.2.3.1 Process discovery

Process discovery is the most common and challenging task in process mining projects to create a process model based on traces captured in the event log. Some algorithms have been proposed for this process discovery task. The models can be presented following notations mentioned in Section 2.2.1 such as transition systems, UML activity diagrams, BPMN, or Petri Nets. These models could then be used for conformance checking, enhancement, and further analysis. The main algorithms for process discovery are:

1) Alpha (α) miner

The α algorithm [75] is one of the first process discovery algorithms that can handle concurrency. This algorithm receives an input of an event log and returns an output of a Place/Transition net (P/T-net). The α algorithm checks the relationships of two activities, e.g. if a task is always followed by another task, it is likely that there is a causal relation between them. The algorithm marks the relationships as: follows (>), causality (→), parallel (||), or choice (#). Those relationships are formalised as follow.

Let W be an event log over T, i.e., $W \subseteq T^*$. Let $a, b \in T$:

1. Follows relation: $a >_w b$ iff there is a trace $\sigma = t_1 t_2 t_3 \ldots t_n$ and $i \in \{1, \ldots, n-1\}$ such that $\sigma \in W$ and $t_1 = a$ and $t_{i+1} = b$,
2. Causality relation: $a \rightarrow_w b$ iff $a >_w b$ and $b \ngtr_w a$,
3. Parallel relation: $a \parallel_w b$ iff $a >_w b$ and $b >_w a$, and
4. Choice relation: $a \#_w b$ iff $a \ngtr_w b$ and $b \ngtr_w$.

These marked relationships are then used to create a process model. The resulting process model contains activities with ingoing arcs and outgoing arcs.
The advantage of the alpha miner is its ability to work on a structured process. The limitation is that it is really depends on the relationship between two tasks. It means that this algorithm cannot correctly mine incomplete and/or noisy event logs and cannot detect the occurrence of duplicated tasks in a process. The alpha algorithm infers wrong relationships in incomplete and/or noisy event logs. Duplicate tasks will never be captured by the alpha algorithm because they will have the same label [76].

2) Fuzzy miner

A fuzzy miner [77] was proposed to overcome problems with unstructured processes. This algorithm uses the concept of *significance* and *correlation* metrics to simplify views of a process at a suitable level of abstraction. Significance measures the relative importance of activities (nodes) and/or their relations (edges). It can be defined based on the level of interest. Correlation measures how two events following one another are closely related. In the simplified model, the algorithm preserves highly significant activities and activity relations (behaviours), aggregates less significant but highly correlated behaviours, and abstracts other behaviours.

The fuzzy miner works based on an approach to measure *long-term relationships*. For example, when the sequence *A,B,C* is found in an event log, the relations are not only $A \rightarrow B$ and $B \rightarrow C$, but also the length-2-relationship $A \rightarrow C$. Subsequently, this algorithm applies three transformation methods to the process model, which are *conflict resolution, edge filtering*, and *aggregation and abstraction*. The conflict resolution solves the problem of conflicting nodes, which include length-2-loops (e.g. $A \rightarrow B \rightarrow A \rightarrow B$), exception (e.g. most of the time $A \rightarrow B$ but there are also insignificant $B \rightarrow A$), and concurrency (i.e. if A and B can be in any order). The edge filtering approach evaluates each edge $A \rightarrow B$ by its utility, which is a weighted sum of its significance and correlation. This is formulated as $util(A,B) = ur.sig(A,B) + (1 - ur).cor(A,B)$, where $ur \in [0,1]$ is a configurable utility ratio. A larger value for *ur* will preserve more significant edges, while a smaller value will only include highly correlated edges. The aggregation and abstraction step preserves highly correlated groups of less-significant nodes as aggregated clusters and removing individual less-significant nodes, based on the *node cut-off* parameter. This approach results in an adjustable level of abstraction of process models.

3) Inductive miner

An inductive miner (IM) [78] is a process discovery framework based on process trees as hierarchical representations of process models. A process tree is an abstract representation of a block-structure network. The IM works by recursively selecting the root operator that best fits the event log, dividing the activities in the log into disjoint sets, and splitting the log using those sets into sub-logs. IM can handle infrequent behaviour and deal with huge models and numbers of event logs. The limitation is that an IM often produces models with high fitness but low precision due to over-generalised behaviours within the log.

There are some variants of Inductive Miner, including the Inductive Miner – infrequent (IMf), Inductive Miner - incompleteness (IMc), and Inductive Miner – life cycle (IMlc). The IMf proposed to add infrequent behaviour filter based on the Pareto principle (80-20 rule) to create an 80% model by filtering out the infrequent activities. Compared to IM, models discovered by IMf have a lower fitness, higher precision, equal generalisation and comparable simplicity [79]. The IMc proposed probabilistic behavioural relations to make IM less sensitive to incompleteness [80]. The IMlc handles life cyle data and distinguishes concurrency and interleaving.

Inductive miner and its variants have also been implemented as a plugin in ProM (see Section 2.2.3.5). The output of the ‘Mine Petri net with inductive miner’ is a Petri net, while the output of the ‘Mine process tree with inductive miner’ is a process tree. A process tree is a hierarchical representation of a process model. The root is ‘seq’ represents sequence/ order of all of its children, with the leaves represent the activities connected by some operators. Those operators are **xor** (one of its children need to be executed), **or** (at least one of its children needs to be executed), **and** (all of its children need to be executed in any order), **concurrent** (all of its children need to be executed and may overlap in time), and **loop** (the first child must be executed and followed by a choice to terminate or execute the second child and the first child, and make the same choice again).

For example, the event log from experiment 1 in case study 1 as discussed in Section 4.3, can be processed using IM and resulted in a process tree or in a statechart as presented in Figure 2.5.

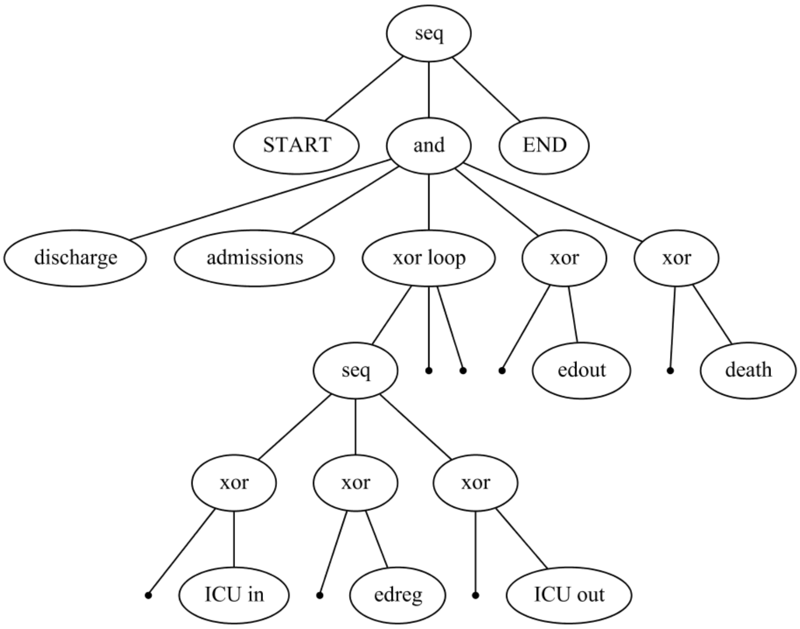


Figure 2.5 An example of a process tree. The root 'seq', with the leaves represent the activities connected by operator **xor, or, and,** or **concurrent.**

The process tree can later be visualised as a process tree itself, as a BPMN, as a state chart, or as a Petri net, using the 'Convert process tree to Petri net' plugin. Inductive miner provides an expressive sematics to create process a model as a process tree, which is convertible into other notations.

4) Heuristics miner

A heuristics miner algorithm [81] focuses on calculating dependency and trace frequencies of events in building a process model. This algorithm consists of three steps: (1) Constructing a dependency graph based on the event log, (2) Establishing the input-output expressions based on the type of dependencies between activities, and (3) Discovering the long-distance dependency relations.

In step 1, a dependency graph is created by analysing causal dependencies. The causal dependencies in a heuristics miner can be seen as an extension of the alpha algorithm. Let W be an event log over T, i.e., $W \subseteq T^*$. Let $a, b \in T$:

1. Directly follows relation: $a >_w b$ iff there is a trace $\sigma = t_1 \, t_2 t_3 \, \dots \, t_n$ and $i \in \{1, \dots, n-1\}$ such that $\sigma \in W$ and $t_1 = a$ and $t_{i+1} = b$,
2. Dependency relation: $a \rightarrow_w b$ iff $a >_w b$ and $b \not>_w a$,
3. Never follow relation: $a \#_w b$ iff $a \not>_w b$ and $b \not>_w a$, and
4. Concurrent relation: $a \parallel_w b$ iff $a >_w b$ and $b >_w a$,

5. Short loop: $a \gg_w b$ iff there is a trace $\sigma = t_1\, t_2 t_3 \ldots t_n$ and $i \in \{1, \ldots, n-2\}$ such that $\sigma \in W$ and $t_1 = a$ and $t_{i+1} =$ and $t_{i+2} = a$,
6. Long distance dependencies: $a >>>_w b$ iff there is a trace $\sigma = t_1\, t_2 t_3 \ldots t_n$ and $i < j$ and $i, j \in \{1, \ldots, n\}$ such that $\sigma \in W$ and $t_1 = a$ and $t_j = b$.

The additional feature proposed in heuristics miner compared to alpha miner is the frequency-based metrics, to indicate the certainty of dependency relation between two activities A and B. In step 2, a causal matrix is then built to map the input-output expressions based on the type of dependencies between activities. In step 3, long distance relationships are considered to be included in the final process model.

The heuristics miner is one of the algorithms with a good performance in process mining [82]. It can be used to discover the main behaviour recorded in an event log. This algorithm can handle noise and incomplete event logs. Heuristics miner has also been implemented as a plugin in ProM as described in Section 2.2.3.5.

Heuristics miner is also the basic algorithm of the interactive Data-aware Heuristics Miner (iDHM) [83], which improved this algorithm with an interactive parameter setting and a built-in conformance checking. This plugin accepts an event log as the input and can produce many results in an interactive manner. Process models can be presented as a directly-follows graph, a dependency graph, a causal net, a data causal net, a data Petri net, and a Petri net. This plugin can export the resulting model for further analysis.

2.2.3.2 Conformance checking

Conformance checking is the second common task in process mining that focuses on checking if the event log conforms to the model and vice versa [4]. The model can be discovered through process mining or created based on the standard expected from the process. This checking can be used to detect, locate, and explain deviations, to quantify trace variants, and to measure performance of the model. When a case in the log does not conform to the model (or the other way around), it can be analysed whether the model does not reflect reality or if the case deviates from the model.

There are four main criteria to evaluate the quality of the discovered model in process mining, which are fitness, precision, generalisation, and simplicity [79, 87]. These four metrics are computed on a scale from 0 to 1, where 1 is optimal. A good model has high values on all four criteria. Each of those will be described as follows.

1) Replay fitness

A model with a good replay fitness allows the behaviour seen in the event log. This means that all traces in the log can be replayed by the model from beginning to end. In other words, a fitness of 1 means that the model can reproduce every trace in the log. There are various ways of defining fitness at the case level or at the event level.

The alignment-based fitness metric [85] is the most commonly used in process mining that compares the sequences of activities in the event log aligned to the process model based on insertions and deletions. The final replay fitness score (Q_{rf}) [86] is calculated as follows:

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

where the denominator is the minimal costs when there is no match between the event log and process model. Conformance checking to check on trace fitness has also been implemented as a plugin in ProM in Section 2.2.3.5.

2) Precision

A model is precise if it does not allow for too much behaviour and it is not underfitting. An underfitting model allows for behaviours very different from what was seen in the log. A precision of 1 indicates that any trace produced by the model is contained in the log.

The alignment-based precision metric [85] calculates based on an aligned event log and compares the number of different activities that occurred to the total number of activities possible in the model. The precision score (Q_p) [85] is calculated as follows:

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

where the context is related to the level of precision being measured, i.e. in the log level or the case level. Conformance checking to check on precision and generalisation has also been implemented as a plugin in ProM as described in Section 2.2.3.5.

3) Generalisation

A model should generalise and not restrict behaviour to the traces seen in the event log [86]. A model that does not generalise is “overfitting”, which means that it specifically fits only the examples in the event log. In a tree representation, consider the frequency of each node to be visited to produce the given log. The formula is as follows:

$$Q_g = 1 - \frac{\Sigma_{nodes}(\sqrt{\#executions})^{-1}}{\#nodes\ in\ tree}$$

The alignment-based generalisation is related to alignment-based fitness and alignment-based precision.

4) Simplicity

This metric is based on the fact that the best model is the simplest model that can explain the behaviour seen in the event log [86]. Simplicity can be measured through complexity, with the lower complexity representing the more simple model. The complexity of a model can be measured by the number of nodes and arcs in the graph. In a tree representation, this can be measured by comparing the size of the tree with the number of activities in the log. If each activity is represented exactly once in the tree, the simplicity is high. This is represented by the following formula.

$$Q_s = 1 - \frac{\#duplicate\ activities + \#missing\ activities}{\#nodes\ in\ process\ tree + \#event\ classes\ in\ event\ log}$$

The effort to make the model simpler is mostly related to pre-processing the steps, for example, by filtering the activities included in the process discovery.

2.2.3.3 Enhancement

The third task in process mining is enhancement. The idea of enhancement is to adapt the target process model to better reflect the reality based on process dicovery and conformance checking tasks. This task repairs or extends the discovered process model using additional information from different perspectives of the process recorded in the event log.

The first type is *process model repair*. A model might need to be repaired to comply with the real execution of the process. For example, if the model shows two sequential activities that in reality can happen in any order, the model may be corrected to reflect the reality. The second type is *the extension*, where additional perspectives are considered to improve the discovered process model, such as the data perspective, the resource perspective, and the performance perspective.

2.2.3.4 Process mining methodology

This section presents several process mining methodologies from the literature. The first formal process mining methodology is the L* life-cycle model, published by the process mining community in the process mining manifesto in 2011 [3]. This model has been further improved as a Process Mining Project Methodology (PM^2) [15], to support iterative analysis. Other methodologies are the Process Diagnostics Method (PDM) [87] to discover several perspectives of a business process and Business Process Analysis in Healthcare environments (BPA-H) [88] for the healthcare domain. Another methodology proposed in healthcare process mining is the Question-Driven Methodology [89] that focuses on the importance of defining questions to be answered with process mining. Related work by Zhou et al. [90] proposed a framework where a business process is continuously optimised using process mining. The ClearPath method [91] extended the PM^2 with a process simulation approach, which is beneficial in engaging with domain experts.

All those methodologies can be related to the general data mining method, such as the Cross-Industry Standard Process for Data Mining (CRISP-DM) [92]. CRISP-DM is the most widely-used analytics model in data mining projects. This model breaks the process of data mining into six major phases: business understanding; data understanding; data preparation; modelling; evaluation; and deployment. An overview of the main process mining methodologies is as follow.

1) L* Life-cycle model

The L* life-cycle model [3] is the first model proposed for process mining projects. This model covers the five stages of a process mining project. The *Planning and justification* (Stage 0) is to understand the data and the domain. The process mining team needs to *extract* the event data, models and other inputs (Stage 1). Those inputs are needed to *create a control-flow model and connect the event log* (Stage 2). When the process is structured, the next step is to *create integrated process model* (Stage 3). The insights can be used for *operational support* (Stage 4).

The L* life-cycle model is useful as a basic sequential methodology. This model provides a general description of each of its stages, making it more flexible in covering different techniques and methods, but the description is not technical enough to be implemented directly. Another limitation is that the L* life-cycle model does not explicitly encourage iterative analysis.

2) Process Mining Project Methodology

The Process Mining Project Methodology (PM2) [15] consists of six stages. The *Planning* (Stage 1) aims to set up the project by identifying research questions, selecting business processes, and composing the project team. The *Extraction* (Stage 2) contains three activities, which are: determining the scope, extracting event data, and transferring process knowledge. The *Data Processing* (Stage 3) is then done by creating views, aggregating events, enriching event logs, and filtering logs. The *Mining and Analysis* (Stage 4) includes process discovery, conformance checking, enhancement, and process analytics. The *Evaluation* (Stage 5) is to diagnose, verify and validate the analysis findings to improvement ideas based on the project goals. The results would be used in *Process Improvement and Support* stage (Stage 6) to modify the actual process execution. The main stages in PM^2 are shown in Figure 2.6.

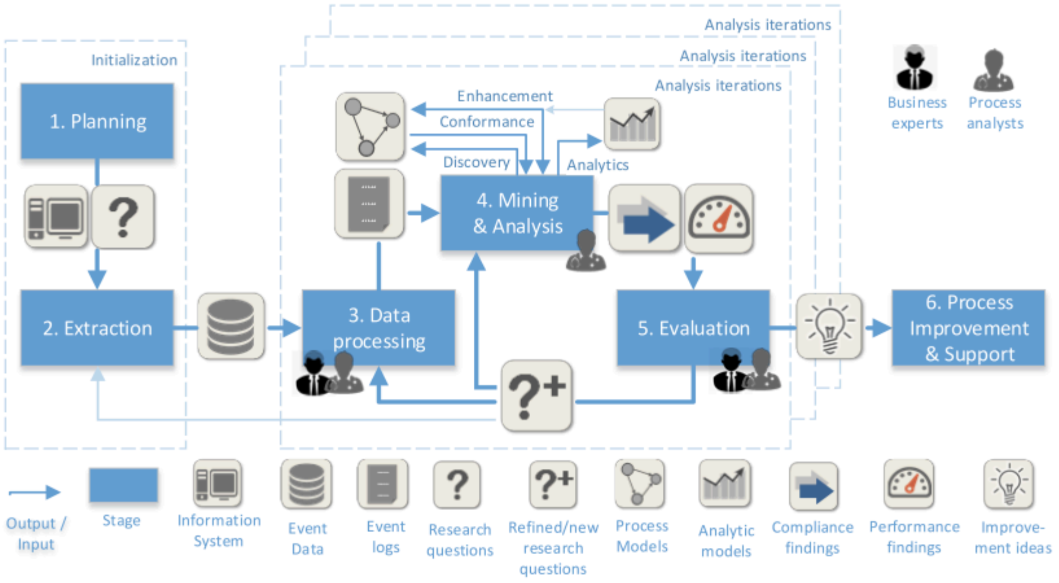


Figure 2.6 The overview of PM^2 methodology. Reproduced from [15]. The key components are listed under them main diagram and the key stakeholders on the top right.

This methodology was designed to support process mining projects aiming to improve process performance or compliance with rules and regulations. This methodology is suitable for both structured and unstructured processes. The PM^2 methodology is highly iterative and emphasises the need for collaboration between process analysts and business experts. It also provides detailed guidance in every stage, which makes it more actionable in real projects. Both L* life-cycle model and the PM^2 methodology are adopted in this study, with more specific approaches needed to adjust to the specific purpose of the experiments.

3) Question-driven methodology

This methodology was proposed by Rojas et al. [89] as an approach for healthcare process mining. Healthcare process mining needs to provide answers to frequently-posed questions about processes in the healthcare system. This methodology contains six stages and each one of those is described in the following paragraph.

The *data extraction* (Stage 1) contains an identification of available data in the Hospital Information System (HIS), ensuring the availability, and verifying the data quality. This includes the identification of frequently-posed questions from the domain experts. The questions drive the steps in the next stages. The *event log creation* (Stage 2) identifies specific data needs, creates the event log, and includes characteristics of each activities. The *filtering* (Stage 3) covers basic, clinical, and question-driven filtering. The *data analysis* (Stage 4) includes the selection of data, statistical, and data mining analyses. The *process mining* (Stage 5) includes identifying the tool, data analysis and process mining cycle. The *results evaluation* (Stage 6) identifies domain experts, defines feedback instruments, and obtains feedback.

4) ClearPath method

This method was proposed by Johnson et al. [91] by extending the PM^2 method with a process simulation approach to address issues of poor quality and missing data. This approach is also useful to support stakeholder engagement. The main stages follow those in PM^2. The main difference is that clinical experts were engaged as interviewees within an iteration and/or in the Clinical Review Board at the end of each iteration. It is suggested that this be done through the use of simulations. In this study, the idea of engaging clinical experts within an iteration is done through discussions providing the visualisations resulted from process mining approaches.

In this research, the L* life-cycle model and the PM^2 method are referred to as two well-known base methodologies. The basic stages are following those of the L* life-cycle model. The iteration approach and detailed steps of the data processing were derived from the PM^2. Other methods including the question-driven method and ClearPath method were used in the detailed steps within the stages. More details about the general methodology followed in this study are described in chapter 3. Some additional methods applied and described in specific experiments.

2.2.3.5 Process mining tools

Process mining tools range from the commercial software to open-source, from implementation of one algorithm to a framework for many algorithms. Some important tools are presented in this section.

1) DISCO

DISCO is a commercial software for process mining produced by Fluxicon [93]. It is a visualisation tool with process models and metrics compatible with ProM. DISCO is focused on being a tool to create visual maps from process data, optimise the performance, control deviations, or explore variations. DISCO is based on the fuzzy miner [77] with several improvements in scalability and robustness. The fuzzy miner allows for simplification and abstraction based on the activities and paths. Further discussion on the fuzzy miner is presented in Section 2.2.3.1.

2) ProM framework

ProM is an open-source process mining framework used for academic purposes that combines different tools and algorithms on the same dataset and compares the mining results [94]. This framework accepts event log input in an XES, MXML, CSV, and a generic eXtensible Markup Language (XML) format, typically contained in an audit trail or transaction log of some complex information systems. Process discovery and conformance checking can be done using several plugins in ProM. The ProM framework allows for interaction between a large number of plugins. A plugin is a module that implements an algorithm, where the implementation agrees with the framework. Some plugins are provided for import, export, mining, analysis, and conversion purposes. Some process discovery plugins used in this study are described in more detail in Section 2.2.3.1.

3) The bupaR packages in R

bupaR is an open-source suite to handle and analyse business process data in R. It was developed by the Business Informatics research group at Hasselt University, Belgium. The bupaR packages provide R libraries to explore and visualise event data and monitoring processes [95]. The main package is bupaR, which provides the basic functionality for handling event data. The other supporting packages are included for exploratory and descriptive analysis, reading and writing eXtensible Event Stream

(XES) files, creating process visualisations, and creating process dashboards. More details about bupaR are given in Appendix F.2.

This study uses ProM as the main tool for process mining because its plugins support many types of analysis in this study. DISCO and bupaR packages are used as supplemental tools. DISCO was used because of the simplicity of use, while bupaR is used because of the availability in R supports a huge range of functionalities.

2.2.4 Process mining in healthcare

Process mining has been useful to analyse healthcare processes for process discovery from event logs [101, 102], for conformance checking [98], and for mapping resources to processes [78, 104]. Previous studies found that process mining algorithms are not sufficiently efficient for unstructured processes [100]. Most mining algorithms have problems in analysing event data from clinical workflows, either due to difficulties in constructing a valid process model or in reflecting reality in the models [73]. Despite the flaws, the concept of process mining carries great potential in helping to analyse clinical workflows and their variations. This sets a strong background to improve currently available process mining techniques for clinical pathway analysis.

A previous review was done by Eric Rojas et al. [101] on process mining in healthcare. They found that most process mining case studies in healthcare were in oncology. However, there were only four different oncology datasets at the time of the review, including the dataset in the Business Process Intelligence Challenge (BPIC) 2011. This finding suggests that there is a great opportunity to find other oncology datasets to be analysed using process mining with a wider range of clinical questions. This review suggested three main algorithms, in process mining for healthcare, which are fuzzy miner [77], heuristics miner [81], and trace clustering [102].

2.2.4.1 Process mining in oncology

As part of this study, a systematic review of previous studies using process mining in oncology has been published [25]. This section summarised that paper.

This systematic review was done in July 2016 to analyse the current literature based on the following query:

> *("process mining" OR "data mining" OR "machine learning" OR "pathway analysis") AND ("event log" OR "patient flow") AND ("oncology" OR "cancer")*

There were 758 papers retrieved in Pubmed, BMJ Open, Journal of Clinical Oncology, ACM DL, and Google Scholar. Three steps were undertaken to find the most related papers: title-based, abstract-based, and full-text filtering. In each step, the article was included if it was: (i) no duplication, (ii) a peer-review conference paper or journal article, and (iii) relevant to process mining in oncology. At the end of the filtering steps, in-depth ancestor search was performed to include articles in the references of the selected papers. As a result of this, 37 papers were selected. Five themes emerged in the study: (1) process and data types; (2) research questions; (3) process mining perspectives, types and tools; (4) methodologies; (5) limitations and future work. A summary of the results is presented in the following paragraphs.

(1) The most commonly used dataset was from the *BPIC 2011* [103], which was used by 24 of the 37 papers. It is an anonymous dataset from the Netherlands which was made available for the challenge. The most common cancer type analysed was *gynaecological cancer* (24 papers), all using the BPIC dataset.

(2) The most common research question was the applicability of process mining in the healthcare domain, specifically in oncology. This research question was broken down into several questions that included *what happened, why did it happen, what will happen,* and *what is the best that can happen.*

(3) All 37 papers applied at least one of three perspectives (control-flow, performance, and organisational) and one of three types (discovery, conformance, and enhancement) of process mining [3]. All papers, except one [104], discussed the control-flow perspective by analysing the pattern of the activity sequences. Most of the papers (27 of the 37) discussed the performance perspective, but only 5 discussed the organisational perspective. All papers, except two [109, 110], studied discovery from a control-flow perspective. In terms of tools, 24 papers used the *ProM toolkit* (www.promtools.org) [94]. ProM toolkit is the de facto standard in the process mining research community and can be combined with other tools, such as R Studio and Java [106], Alchemy and BUSL [107]. Other papers proposed their own tool [113–120]. In case studies other than oncology, process mining was implemented using the DISCO commercial tool (www.fluxicon.com/disco), such as in [121–123].

(4) Only one paper [119] clearly mentioned the *L* life-cycle model* as the methodology used in the study. Eleven papers [84, 110, 112, 114, 125–132] proposed new algorithms and/or techniques. The methodology followed by the other ten papers was using available plugins and/or functionalities in existing tools to solve the problem in their case study.

(5) The paper identified data, techniques, and team limitations. Data limitations were related to limited access to the data, data quality problems, attributes not available from the data being extracted, or the dataset was available in inappropriate levels of detail [108, 114, 115, 125–127, 134–136]. Technique limitations were related to the chosen functionalities. Team limitations were identified in two papers [136, 137].

This systematic review gave insight that process mining is applicable in oncology, and there is a great opportunity for this to be improved,especially with regard to the technical aspect. This research works on the control-flow and performance perspectives. In terms of process mining types, this research will explore discovery, conformance checking, and enhancement. The literature review suggested that the most widely-used package is the ProM toolkit, which will also be used in this research.

2.2.4.2 Challenges for process mining in healthcare

An obvious challenge for process mining in healthcare is the data quality. Process mining projects work with event logs, which are automatically generated by the information system within a hospital. Data quality issues in process mining as discussed in a book by van der Aalst [4] can be related to the quality of the event logs.

In this study, the quality of data being used in process mining was assessed using Weiskopf and Weng framework [133] as a generic data quality assessment approach. This framework structures data quality assessment in five dimensions and seven methods. The five dimensions are completeness, correctness, concordance, plausibility, and currency. The seven methods are comparison with gold standards, data element agreement, data source agreement, distribution comparison, validity checks, log review, and element presence.

A previous study by Homayounfar [134] found that the challenging characteristics of healthcare processes are that they are complex, multidisciplinary, ad hoc, and dynamic. The *complexity* is mainly caused by the heterogeneity of the diseases, the treatments of patients, and the clinical expert judgements. This is also related to the *multidisciplinary* characteristic of healthcare processes. Treatment of a patient would involve hospital departments consisting of many different roles (doctors, nurses, etc.) which are highly specialised in their areas.

The main challenge addressed in this study is related to the *ad-hoc* and *dynamic* nature of the healthcare processes. *Ad hoc* changes in healthcare processes are an inevitable result of the *dynamic* nature of the healthcare provision. Those changes can happen within different levels of the healthcare system, for example, a new clinical target from the government, an introduction of new procedures, or technological developments.

This is related to Leavitt's diamond [135], where the healthcare data can be seen as an output of a complex relation of four forces (structure, process, technology, people). The idea is that in any organisation, everything is connected and changing one thing can impact another. Those four forces are always changing and affecting each other. For example, a change in the technology used in the healthcare information system will change the way a task is done by the people in the hospital organisational structure. It is therefore not suitable to treat the healthcare process as a static process and analysing the changes over time becomes crucial. The four interconnected forces of Leavitt's diamond are illustrated in Figure 2.7.

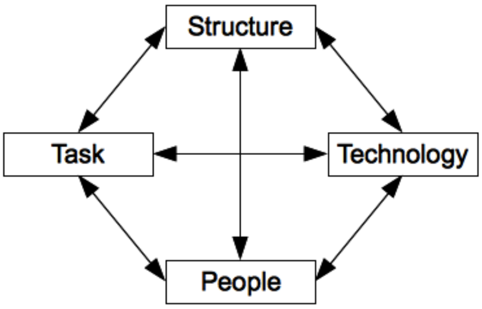


Figure 2.7 Leavitt's diamond [135]. It illustrates the four interconnected forces (structure, process, technology, people) in healthcare data.

A common analysis of healthcare processes by treating dataset across a long duration of time as one static dataset would result in a high number of variants, which is difficult to interpret. The dynamic characteristic is especially interesting in this study, which further analyses the healthcare process changes over time. Those four forces are being analysed through discussions with both clinical- and technical experts.

2.2.5 Process mining to analyse process changes

2.2.5.1 Concept drift

In process analytics, the process might be changing while being analysed, due to periodic/seasonal changes, or due to changing conditions. This condition is known as *concept drift* [136]. Two types of concept drift detection approaches have been used, i.e. stream evolution monitoring [142, 143] and data distribution comparison in two time windows [144, 145].

Three challenges exist when dealing with concept drift. These are:

1) Change point detection (Did the process change? If so, when?)
2) Change localisation and characterisation (What has changed?)
3) Change process discovery (How to unravel the process change?)

There are four perspectives in business process analysis: *control flow, data, resource,* and *time perspectives*. One or more of these perspectives may change over time. The control flow perspective deals with the behavioural and structural changes in a process model. The data perspective refers to the changes in the requirement, usage, and generation of data in a process. The resource perspective deals with the changes in resources, their roles, and organisational structure, in relation to the process. The time perspective concerns the timing and frequency of events [141].

Four classes of drift are: *sudden drift, gradual drift, recurring drift,* and *incremental drift*. Sudden drift refers to a substitution of an existing process P_1 with a new process P_2 where P_1 ceases to exist from the moment of substitution. Gradual drift refers to a scenario where a current process P_1 is replaced with a new process P_2 where both processes coexist for some time with P_1 discontinued gradually. Recurring drift corresponds to a scenario where a set of processes P_1 and P_N reappear after some time (substituted back and forth) that commonly caused by a seasonal influence. Incremental drift refers to a scenario where a substitution of process P_1 with P_N is effected via smaller incremental changes [16].

Most approaches use a sliding window approach [146–148], where events in different windows are compared using statistical methods to detect significant changes. The data of process features are split into groups based on time windows. The options are to split it with overlapping or non-overlapping windows, with same-size or same-duration windows, or with a more advanced technique of adaptive windowing. For each iteration, two subsequent windows are compared statistically to investigate if there is a significant difference between the two windows. More details about the statistical approach are presented in Section 2.2.6.

A new approach is proposed in this study to use process mining techniques to detect, localise and characterise process change over time. The main input of this approach is an event log. An event log L can be split into sub-logs of s traces each. An additional challenge is that an event log consists of traces that span in a time duration rather than data points. More details about the statistical approach are presented in Section 2.2.6.

2.2.5.2 Change point detection

In process change analysis, the first challenge is to detect concept drift in the processes and to identify the periods at which these changes have taken place. Change point detection involves two primary steps, which are capturing the characteristics of the traces, and identifying when the characteristics change. Some approaches to detect a change point include statistical testing [141], trace clustering [144], or abstract representation [145].

A change point can be detected by *statistical testing* over feature vectors. Potential control-flow changes in the processes over time are detected through analysing the event log. Statistical hypothesis testing is used to evaluate and compare groups of data. Because there is no known a priori distribution of the feature values in an event log, non-parametric tests are suitable methods for change point detection. For each iteration, two subsequent sub-logs are compared, which means that two-sample tests are needed. Both univariate and multivariate hypothesis tests are considered, because the comparison can be made to one activity or a set of activities. The limitation is that it requires identification of the features and window sizes for change point detection [141].

Another approach for change point detection is *trace clustering*. The idea of change point detection using trace clustering is to cluster the traces inside a time window based on the average distance between each pair of activities in the traces. A similarity matrix is used to record the similarity between cases. To detect potentially interesting change points, compute the change in the values of the similarity matrix over time. This method requires that window size be defined; it also cannot deal with loops [144].

Another approach is based on an abstract representation of a *polyhedron*. This approach sets prefixes in a random sample of traces in the event log and computes the fitness of subsequent prefixes of traces against the constructed polyhedron. A polyhedron can be described as the sets of solutions of a set of linear inequality constraints with rational (Q) coefficients. Let P be a polyhedron over Q^n, then it can be represented as the solution to some system of m inequalities $P = \{X|AX \leq B\}$ where $A \in Q^{m \times n}$ and $B \in Q^m$. The domain of polyhedra provides the operations required in abstract interpretation, including intersection and join. The limitation of this approach is that the entire detection process has to be executed from the start, which decreases the scalability of the approach [145].

2.2.5.3 Change localisation and characterisation

When a process change has been detected, the next step is to localise the regions of change and to characterise the nature of the change. This is related to both the nature of the change (sudden, seasonal, gradual, or incremental) and the perspective of the change (control-flow, data, resource, or performance). The general approach [16] is to analyse each activity pair individually or as a subset. One important task in the analysis of process change is process comparison. Four papers proposed several aspects related to process comparison are presented in the following paragraphs.

Delta analysis [146] provides a basis for process comparison by generating a similarity measure between the model and event logs. This method maps sequences of the traces, compares them to the reference model, and evolves the reference model based on the deviations in the log. This method is suitable for business process improvement in general, but allows free modelling of activity flow that result in an exponential space and time to check whether a log fits a process model.

Another approach has been to explore *three similarity metrics* [147], including (i) node matching similarity that compares the labels and attributes of the process model elements, (ii) structural similarity that compares element labels as well as the topology of process models, and (iii) behavioural similarity that compares element labels as well as causal relations captured in the process model. The node matching similarity was based on pairwise comparisons of nodes or attributes, the structural similarity was based on the graph-edit distance between two graphs, and the behavioural similarity was based on the indirect relations between nodes or attributes. This approach is simple as it is based on the well-known causality graph, but the experimental results showed that the time performance is not optimal.

A cross-organisational comparison [118] using comparison points with process mining techniques has also been proposed. Using this approach, the models of a similar process in many different organisations can be compared based on pre-defined comparison points. This approach is limited as it relies on the domain experts to provide a set of pre-defined comparison points.

In general, there are four categories of process comparison, which are: model-based comparison, conformance-based comparison, log-based comparison, and performance-based comparison. *Model-based comparison* is based on the control-flow comparison, where the structural properties of the models are compared [148]. A limitation of this approach is that the differences in terms of frequency or any other process metrics are not detectable. *Conformance-based comparison* is based on the conformance perspective of process mining. Two event logs can be compared based on their conformance to one reference model. A limitation of this approach is that it relies on the validity of the reference model. *Log-based comparison* [149] detects relevant differences between processes based on the event logs, as implemented in the 'Process Comparator' plugin in ProM [150]. This approach takes two event logs and visualises the differences using annotated transition systems. A limitation of this approach is that the event logs need to be completely representative of the process. *Performance-based comparison* compares process performance based on pre-defined comparison points such as waiting times, throughput, and Length of Stay (LoS) [118]. A limitation of this approach is that it relies on the availability and capability of domain experts to define the comparison points.

2.2.5.4 Change process discovery

The next step is to discover process change or to unravel the nature of the change. One important approach for change process discovery is through focus group discussions with domain experts [151] to discuss the findings and reveal the nature of change. This method involves a group of people participating in an interactive discussion focussing on specific issues. The nature of change can be related to one or more forces in the Leavitt's diamond, i.e. the technology, people, structure, or task/ process. Process mining approaches can be used to support the discovery of process change based on the recorded data of the process in the event log.

In terms of process change analysis, this study focuses on an offline setting for process change analysis in healthcare. The change point detection is done using a windowing approach to find significant changes in the execution of the process over time. The statistical approach is used to test the hypothesis of this study. Process comparison is done based on all four categories, which are model-based, log-based, conformance-based, and performance-based comparisons. The purpose of using all of those comparison categories is to get a thorough understanding of the process change.

2.2.5.5 Process mining and user interface (UI) design

In this research, process mining is used to test the effects of UI design on care processes within an EHR system. The goal of UI design is to create systems that are modelled based on the characteristics and tasks of the users. The systems are built to increase user productivity, satisfaction, and acceptance as well as decrease user errors, and user training time [152].

An EHR system is an information system processing and managing the patient records in a hospital. It covers organisational processes such as medical order entry and the medical treatment processes such as diagnostic procedures for a particular patient [153]. While organisational processes help to coordinate healthcare professionals and organisational units, treatment processes are linked to the patient. Those two types of processes should be covered within a hospital, either separately or in a centralised system, to support better care delivery to the patients.

Healthcare providers are challenged by the increasing amount of information collected routinely in clinical settings. There is a greater need to utilise technologies to manage such information efficiently. Information technology is changing the way patient information is obtained and gathered and can impact the decision-making processes of clinicians. For example, if poor information is displayed, the delivery of care might be inefficient, which may include redundant ordering of tests or missing important information in the diagnosis of the patient. The key is to have the right information at the right time in the right place for the right clinicians.

The previous literature summarised in this section builds an understanding of the challenges of this study. There are some techniques implemented in the literature, including process change analysis and concept drift, but they haven’t been implemented in the healthcare domain. As presented in Section 2.2.4.2, process mining in healthcare is challenging, with one particular challenge on the change over time. In this research, process change analysis will focus on how EHR system change affects the actual process represented in the event logs of the EHR system.

2.2.6 Statistical approach

Statistical approaches have been used in process mining projects. Different statistical approaches can be used in descriptive and inferential ways to describe the event log [86, 106], to check conformance of the model to the log [153], or to compare between two logs [16]. These references show that the statistical approach can be applied in different stages of process mining projects to improve the confidence of the findings. In this study, descriptive statistics are used to describe the sample data being collected, while the inferential statistics of the sample data are used in the process comparison.

The main statistical tests are required in the process comparison step of this study. Hypothesis testing was needed to evaluate and compare groups of data. The choice of a particular test is dependent on the nature of the data and the objectives of the experiment. *The parametric test* can be used when the data have a particular distribution, e.g., normal distribution, and *non-parametric test* if no particular distribution is known. There are two groups of data to be compared, so that the suitable test is a *two-sample test*. Tests dealing with scalar data elements are called *univariate tests*, while those dealing with vector data elements are *multivariate tests*. For example, the *independent univariate two-sample t-tests* are suitable where data are collected from two separate groups of parametric scalar data [154].

There are some plugins in ProM that implement statistical tests, including Process Comparator and Concept Drift plugins. These two plugins were explored in the early stage of the study. Both of them are not sufficient to be used as the only technique in this study and need to be combined with other techniques. These two plugins are described in the following subsections.

2.2.6.1 Process comparator: compare variants of a process

The Process Comparator plugin [150] compares variants of a process, which can be derived from the same process in different locations, in different groups of people, or in different timeframes. This plugin requires two event logs or more to be compared. The differences visualised using transition systems annotated with measurements.

The statistical test used in this plugin is the two-tailed “Welch’s t-test”, also known as the “two-tailed t-test with different variances”. This test is suitable when the two sets of measurements come from independent populations, such as from two event logs from two groups of patients. The results are visualised in an annotated transition system, using visual properties (e.g. thickness and colour) of nodes and arcs. The graphic visualisation can be adjusted to represent trace frequency, elapsed time, sojourn time, remaining time, or duration. This plugin also allows us to filter out rare behaviour with the frequency filtering capability. The default alpha significance level (α) in this test is 5%.

The differences will be presented in a range of colours based on the effect size oracle which, given two multisets of measurements, returns the size of the effect (i.e. how small or large is the difference) and the sign of the difference (+/-) within a certain scale. Cohen’s *d* is used to measure effect size, which measures the difference of sample means in terms of pooled standard deviation units. The ranges of *d* values can be categorised: $d = \pm 0.2$ is considered as a small effect, $d = \pm 0.5$ is considered as a medium effect and $d = \pm 0.8$ is considered as a large effect.

2.2.6.2 Concept drift: hypothesis test to analyse process changes

The Concept Drift plugin [16] is available in ProM for hypothesis testing and is useful in analysing process changes. The analysis of the process change was done by splitting the event log into two sub-logs (before-log and after-log) with non-overlapping windows. The change point can be set to a specific date where there may be a specific change that has happened on a known date. In order to make this test more general, the date of the change point will be incrementally changed with a moving window approach. The test would then compare those two logs based on several comparison methods. Statistical hypothesis testing can then be used to evaluate the differences between the before- and after-logs.

Based on general statistical approaches, two-sample univariate and multivariate tests are used. The two hypothesis tests for univariate two-sample are Kolmogorov-Smirnov test (KS test) and the Mann-Whitney U test (MW test), and the test for multivariate data is the two-sample Hotelling T^2 test. The hypothesis tested by the KS test is “Do the two independent samples (populations P1 and P2) represent two different cumulative frequency distributions?”, while the hypothesis tested by the MW test is “Do the two independent samples have different distributions with respect to the rank-ordering of the values?”. The multivariate Hotelling T^2 test is a generalisation of the t-test and evaluates the hypothesis “Do the two samples have the same mean pattern?”. All of these tests yield a significance probability assessing the validity of the hypothesis on the samples [141].

The suggested framework for analysing concept drifts in process mining consists of five steps: (1) feature extraction and selection, (2) generate populations, (3) compare populations, (4) interactive visualisation, and (5) analyse changes. The main input is an event log, and the main output is the detected change.

2.3 Summary

This chapter has reviewed the related literature that builds the background of this study, including the published systematic literature review [25] in 2016 that has been cited in 28 articles. The complexity of the healthcare domain makes it important to build an understanding of healthcare background. Cancer as a specific disease analysed in this study to understand the definition, characteristics, and the target of cancer treatment. The technical background in this study is in process mining and how this approach has been applied in healthcare and process change analysis.

Based on the literature review on the healthcare background, it is evidenced that there are limited numbers of process mining projects in healthcare. Most process mining studies in healthcare have used artificial data provided in process mining challenges. On the other side, with the high complexity nature of healthcare data, process mining is a potential approach to analyse the processes in healthcare settings. One of the important challenges is that many aspects of healthcare practices might change over time. Based on the literature of the technical background, it is shown that many articles have been published to propose different techniques for process mining and detecting change from the data, but none of them have been specifically proposed to tackle the complexity of healthcare data. Based on this understanding of the current literature, the general methodology and data sources to be used in this study are presented and described in Chapter 3.

Chapter 3
Methodology

Chapter 2 presented the healthcare and technical background that helped build an understanding of the literature for this study. Chapter 3 presents the general methodology, the details of each stage and the three datasets that have been used in this study. The general methodology has been built based on the background knowledge of the complex nature of healthcare data and the suitability of the available techniques in process mining and process change analysis in the literature. The general methodology has been adjusted to the specific characteristics of each case study. One additional methodology related to this chapter has been presented and published as a jointly-authored publication in the 2018 Process-Oriented Data Science for Healthcare entitled “The ClearPath method for care pathway process mining and simulation” [91].

3.1 Main stages of the general methodology

The main methodology used in this research is the Process Mining Project Methodology (PM^2) [15] as an improvement of the original L* life-cycle model [3]. Table 3.1 shows the main stages in those two methods, along with the main stages of the general methodology in this study.

Table 3.1 Main stage development

Process mining method		
L* life-cycle (2011)	**PM^2 (2015)**	**This study**
0. Plan and justify	1. Planning	1. Planning and justification
1. Extract	2. Extraction	2. Extraction, Transformation, and Loading (ETL)
	3. Data processing	
2. Create control-flow model and connect event log	4. Mining and analysis	3. Mining and analysis
3. Create integrated process model		
<< side stage during Stages 2 to 4: interpret, redesign, adjust, intervene, support >>	5. Evaluation	4. Evaluation
4. Operational support	6. Process improvement and support	<< not applicable >>

As presented in Table 3.1, the methodology in this study commenced with *Stage 1 (Planning and justification*). Both the L* life-cycle model and the PM^2 method started with the planning stage, to set the scene and justify the importance of the process mining project. In the L* life-cycle model, Stage 1 consists of understanding the available data and understanding of the domain, while the PM^2 method suggested that Stage 1 consists of selecting the business process, identifying research questions, and composing the project team. In this study, Stage 1 followed the same stage as in the PM^2 method but extended it with the question-driven method for identifying research questions [89]. It is also important to note that very little is mentioned about justification in the PM^2 methodology. In this study, justification is important as this is necessary for academic research to justify the planning scientifically. The suitability of the data to be used in this study has also been extensively explored through a data quality assessment following the Weiskopf & Weng framework [133].

Stage 2 (ETL) is the next stage of this study. It combines the extraction and data processing stages in the PM^2 method. It is renamed to make it consistent with the general data analysis approach, i.e. ETL. This was also done to reflect the specific conditions in the three datasets in this study. The full datasets were accessible for this study, and each analysis required an iteration of ETL as a subset of the datasets based on a specific cohort of patients. The L* life-cycle model did not explain the extraction steps in detail but suggested that the output of this stage are historical data, handmade models, objectives, and questions. The PM^2 method specified three steps in the extraction steps: determining scope, extracting event data, and transferring process knowledge. The PM^2 method explicitly added a data processing stage that consisted of creating views, aggregating events, enriching logs, and filtering logs. In this study, both extraction and data processing were combined and renamed as *Stage 2 (ETL)*. The name was chosen to improve clarity and was based on the basic approach for general data analysis. The steps in Stage 2 follow the steps suggested in the extraction and data processing stages in the PM^2 method.

Stage 3 (Mining and analysis) was named after the same stage in the PM^2 method and is the main part of this study. In the L* life-cycle model, process discovery and conformance checking are done in the 'create control-flow model and connect event log', while the enhancement is done in the 'create integrated process model'. This study followed the PM^2 method, where process mining and process analytics are combined into one stage. Process mining includes process discovery, conformance checking, and enhancement. Process analytics were focused on process change analysis based on a concept drift analysis [16]. Process analytics in the Patient Pathway Manager (PPM) Cancer case study was done with and without a known change. When a change is known, the event log is split into before and after the change. It is followed by process comparison of those two sub-logs. When there is no prior information on a change, other methods adopted in the process analytics are the signal decomposition method [155] and the Statistical Process Control (SPC) method [156].

Stage 4 (Evaluation) was the final stage of this study. In the L* life-cycle model, the evaluation was included as additional steps to interpret, redesign, adjust, intervene, and support the analysis of the process of interest. In the PM^2 method, the evaluation was done to diagnose, verify, and validate the results of the previous stages. In this study, Stage 4 was done to evaluate the findings in the statistical and clinical perspectives. The statistical evaluation done in this study was focused on the hypothesis testing to find statistically significant differences between the sub-logs over time. The clinical evaluation was done to ensure that the results are meaningful from a clinical perspective.

The methods were followed by the *Operational support* in the L* life-cycle model and by the *Process improvement and support* in the PM^2 method. Both stages suggested the implementation of the findings into the real-life process. This is not applicable in this study and is not included in the method.

3.2 General methodology

The general research methodology was developed as described in the previous section. It is based on the PM^2 method and is extended for process change analysis. There are four main stages as mentioned in Section 3.1: (1) Planning and justification, (2) ETL, (3) Mining and analysis, and (4) Evaluation. The overview of the general methodology is presented in Figure 3.1.

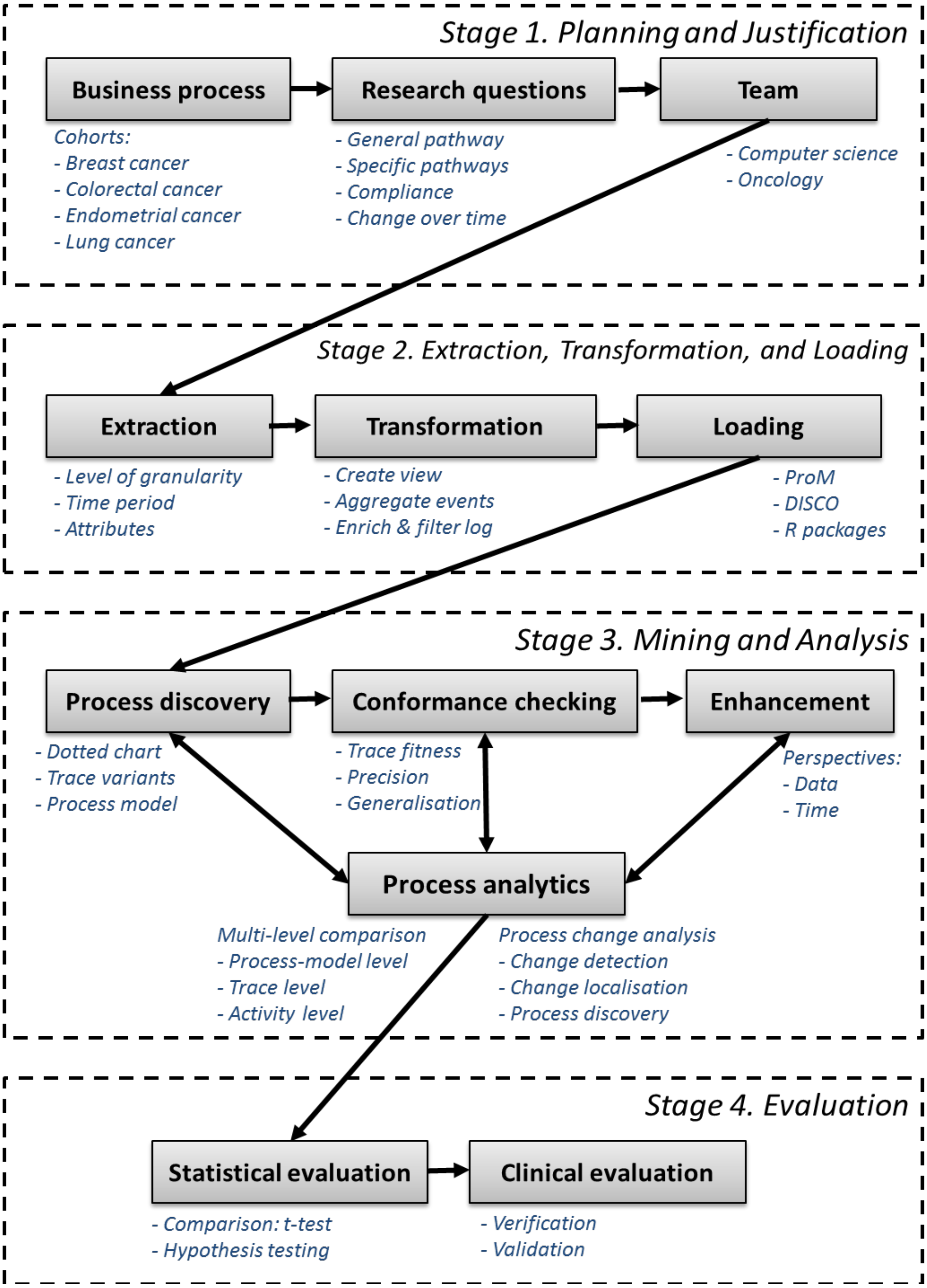


Figure 3.1 Research methodology. The dashed blocks represent main stages, the boxes represent steps within a stage, connected by straight arrows. The blue texts explain the details of each step specific in this study.

Each stage presented in Figure 3.1 is discussed in more detail in Sections 3.2.1–3.2.4. Process change analysis is an important part of this study and is described separately in Section 3.3.

3.2.1 Planning and justification

This stage focused on identifying the business process, research questions, and project team as starting points for the project. This stage followed PM^2 except on: (i) additional data quality assessment which followed the Weiskopf & Weng framework, (ii) research questions were identified which followed the question-based methodology, and (iii) project team involvement followed the ClearPath method.

In this study, the business process was defined based on a cohort of cancer patients, which included colorectal cancer, breast cancer, and endometrial cancer patients. Those cohorts were chosen through careful discussion with clinical experts. A careful data quality assessment was also done to check if the dataset was suitable for process mining. The assessment was done following the Weiskopf & Weng framework. Datasets were assessed to check their completeness, correctness, concordance, plausibility, and currency to be analysed with process mining.

Research questions were identified based on previous related studies and general research questions for process mining projects. The additional methodology referred to in this stage was question-driven methodology, which suggests that frequently posed questions are identified from the very beginning of the data analysis. Some frequently posed questions were adapted as the initial questions:

1) What are the most followed paths and the exceptional paths?
2) Are there differences in care paths followed by different patient groups?
3) Do we comply with internal and external guidelines?
4) How was the process changed over time?
5) Can process mining be used to analyse the effect of User Interface (UI) changes to the care process?

Project team identification was defined in the early stage of this study, which included computer scientists and clinical experts. The two clinical experts are: (1) Professor Geoff Hall, a senior oncologist and senior lecturer in Medical Oncology and Cancer Informatics, and (2) Dr Kieran Zucker, an honorary clinical oncology registrar and clinical research fellow at the University of Leeds. Based on the ClearPath method, an important focus on composing the project team was to engage clinical experts through a clinical review meeting at the end of every stage of the study.

3.2.2 Extraction, transformation, and loading

This stage follows the PM^2 method except that the full datasets were accessible so that a sequence of ETL was done for each experiment. The additional step is to initially re-build the database in a database management system or to access the database directly.

The ***Extraction*** stage was done to extract event data and (whenever possible) reference models from the system, based on the initial planning and justification made in the previous stage. The extraction scope can be defined for each cohort of patients, based on the focus on the analysis in that cohort. This stage requires an understanding of the available data to determine the level of granularity, time period, and attributes needed for the analysis.

For formalisation purpose in this study, the following definitions were used:

> **Definition 1 (Event logs)**. An event log ***E*** is a set of events *(**c**, **a**, **t**)*. An event happening in a timestamp ***t*** is described by a case identifier ***c*** and an activity label ***a***. A trace ***T*** is a sequence of a subset of events happening to a case ***c*** ordered by a timestamp ***t**, where* $T \in E$.
>
> **Definition 2 (Process models)**. A process model ***M*** is a directed graph modelling the traces ***T*** in the event log ***E***. The process model ***M*** draws activities ***a*** as nodes and the possible paths ***p*** between nodes as arcs from one node to another. Standard process mining algorithms can be used to discover process models with additional components identified, such as the frequency of nodes and arcs as the occurrence of ***a*** and ***p*** in ***E***, respectively.

The ***Transformation*** stage was based on the data processing stage in PM^2, as follows:

1) Creating views

The views created in this study are based on the understanding of the data structure and the research questions. Generally, the view would be to analyse the patient journey during a cancer treatment pathway, where patient ID is the case ID, a particular action during the treatment is the activity, the clinician who did the activity is the resource, and the time when that activity was done by the resource is the timestamp. The event log is then a collection of events recorded in {*case ID, activity, resources, timestamp*} format.

2) Aggregating events

The events might need to be aggregated to reduce complexity and improve the readability of the process model being discovered. This can be done to get process models with different levels of details, based on the details required in the research question.

An example in this study is in the analysis of a chemotherapy pathway. In the analysis of chemotherapy cycles, fine-grained event names were used, which are *Cycle 1, Cycle 2, Cycle 3*, etc. When it was required to get a general pathway of patient journeys, the event name is *Chemotherapy*, which was then being analysed with other events such as *Referral, Outpatient,* and *Surgery*.

3) Enriching logs

Log enrichment was done by adding information to the event log. This can be done by computing additional data or by adding external data. An example in this study was by adding duration and year of diagnosis. The process duration for each patient was calculated as the number of days between the first activity and the last activity. The year of diagnosis was used to group the patients based on the year of diagnosis date for each patient.

4) Filtering logs

The last type of transformation is the filtering, which can be done based on the attribute, the variance, and/or the compliance. Attribute-based filtering was done by removing events or traces based on the values of a specific attribute, such as the event names, or the duration. For example, in one experiment of the PPM Cancer case study, to focus on the pathways from referral to diagnosis, all events that happened before referral and after diagnosis would be filtered out. Variance-based filtering was done to group similar traces to split the event log to discover simpler process models. For example, in this study, the log was partitioned based on the year of the diagnosis. Compliance-based filtering was done to remove traces or events that do not comply with a given rule or process model. For example, in this study, patients were not included if there was no referral recorded in the 120 days before a cancer diagnosis.

The ***Loading*** stage was done by loading the extracted and transformed event data to the process mining tools, including DISCO [93], ProM [94], and bupaR [95]. In DISCO and ProM, extraction and transformation are straight forward and are easily performed through the UI. A form is provided in both tools to upload an event log in .csv or .xes file, then specify the case-id, event-id, resource, start- and end- time. In bupaR, an event log can be recorded as a data frame, for example, to create an event log from a table {*PID, act, time*}, as follow:

```
event_log %>%
  eventlog(case_id = "PID", activity_id = "act", timestamp = "time")
```

DISCO was used to gain a quick insight from an event log along with the statistical details of the event log for further analysis. The limitation is that DISCO only provided a process discovery using a fuzzy miner. ProM was used to explore other algorithms for process discovery, such as a heuristics miner [83] and an inductive miner [157]. The bupaR was used for better support in statistical analysis and discussion with clinical experts.

3.2.3 Mining and analysis

This stage includes the main tasks of process mining, which are: process discovery, conformance checking, and enhancement. Those were done as a process analytics approach to gain insights about the process from the data. Each of those main tasks is described in the following sections.

3.2.3.1 Process discovery

In this study, process discovery was performed using several available algorithms DISCO, ProM, or R language. The main algorithms used for process discovery are the fuzzy miner in DISCO, the interactive Data-aware Heuristics Miner (iDHM) plugin in ProM 6.8, and bupaR in the R language. The fuzzy miner in DISCO was mainly used in the MIMIC-III case study, because it was straight forward and reliable to use in common situations. The iDHM plugin in ProM was used as a comparable tool that provided better connectivity with other tasks in ProM through data filtering and conformance checking plugins. Especially in the PPM Cancer case study, bupaR was used because the R language was originally supported in the hospital-networked PC, while DISCO and ProM had limited support due to regular updates that needed to be done by IT support in the hospital. More details about fuzzy miner, heuristics miner, and bupaR have been presented in Section 2.2.3.1. The decision to use an algorithm and a tool was based on the data characteristics, the capability of the algorithm to handle that type of data, and the suitability of the output of the algorithm for the next analysis in the study.

In this study, the results of process discovery task were presented through process models, trace variants, dotted charts, and other visualisations from the event log. The main visualisation is a process model, which usually followed by other visualisations to support the specific analysis. Unless otherwise stated, the properties of each visualisation are described as follows.

1) Process model

A process model presents a sequence of events as defined in Definition 2 in Section 3.2.2. In this study, a process model can be visualised as a directly-follows graph, a heuristics net, a state transition diagram or a Petri Net.

Figure 3.2 shows an example of a process model from an experiment in Chapter 6. It begins with a START node and finishes with an END node. These properties of the process model resulted from three tools DISCO, bupaR, and ProM.

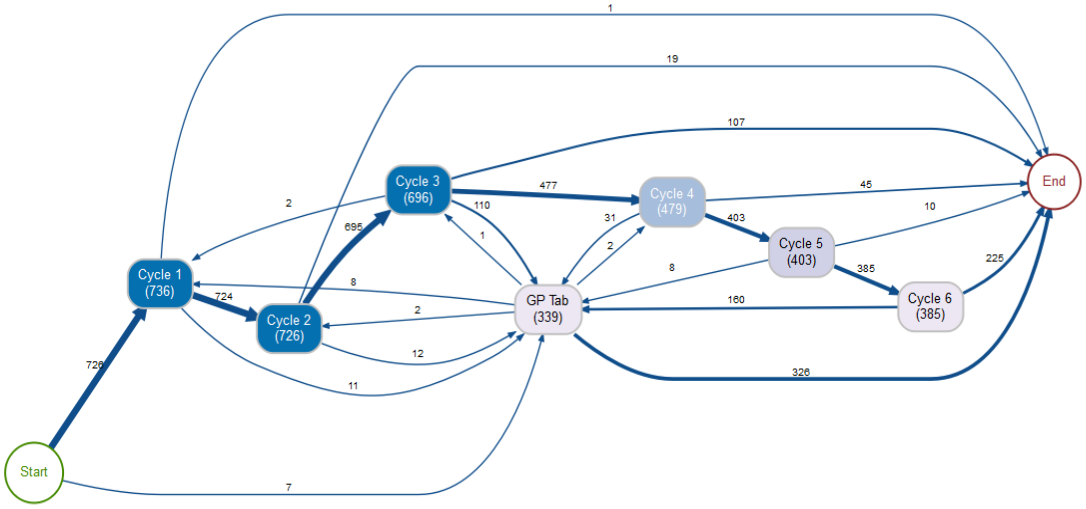


Figure 3.2 An example of a process model (bupaR). A node represents an activity and an arc represents a path from one activity to another. Nodes and arcs are labelled with the number of patients having those activities and paths. Nodes and arcs are colour-coded with a darker colour representing a more frequent activity or path.

The input is event log in .csv or .xes format. In ProM, an event log in .csv format can be transformed into .xes format using a plugin “Convert CSV to XES”. The event log can go through some transformation steps as described in Section 3.2.2. The tools would then create a process model. In DISCO, the process model can be seen in the

"*Map*" tab and is adjustable to filter based on the percentage of the most frequent activities and/or paths. In bupaR, the syntax to create a process model from an event log is as follow:

```
event_log %>% process_map()
```

There are also some other syntaxes to adjust the frequency value shown in the process model. For example, to show *relative frequency* and *median (day)*, as follow:

```
event_log %>% process_map(type=frequency("relative"))
event_log %>% process_map(performance(median,"day"))
```

In ProM, an additional step is needed to select a process discovery plugin, for example by choosing a plugin "interactive Data-aware Heuristics Miner (iDHM)". The resulting process model could then be adjusted to show some features needed for the next analysis.

2) Trace variant diagram

A trace variant diagram represents trace variants as the sequence of events. A trace variant diagram can be created in ProM, DISCO, or bupaR. Figure 3.3 shows an example of a trace variant diagram from an experiment in Chapter 6, showing the most frequent trace variants (top five variants showing >85%).

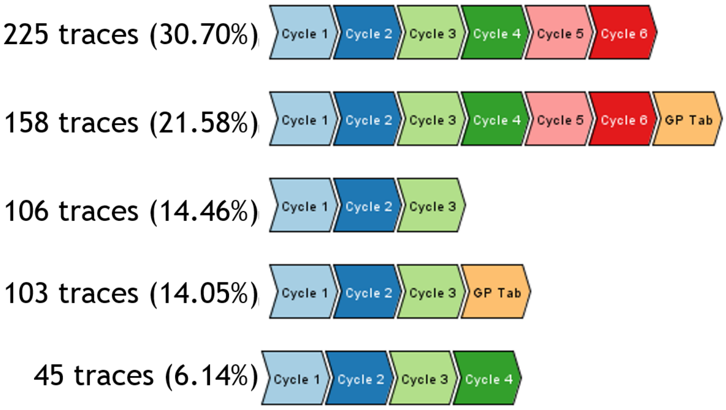


Figure 3.3 An example of a trace variant (ProM). A colour-coded and named shape represents an event. Each line represents one trace variant. Additional information on the left show the number of traces of each trace variant and the percentage from the complete log.

In ProM 6.8, an event log can be visualised in a dotted chart by *Select visualisation >Explore Event Log (Trace Variants/ Searchable/ Sortable) (LogEnhancement)*. In DISCO, trace variants can be found in the *Cases* tab and are shown as a table or a flowchart-like graph. In bupaR, trace variants can be shown using following syntaxes:

```
event_log %>% trace_explorer()
event_log %>% trace_explorer(coverage=0.8)
```

The two syntaxes show all trace variants and the top 80% variants, respectively.

3) Dotted chart

A dotted chart shows traces over time. In this research, a dotted chart shows patient pathways over the treatment duration. Figure 3.4 shows an example of a dotted chart from an experiment in Chapter 5.

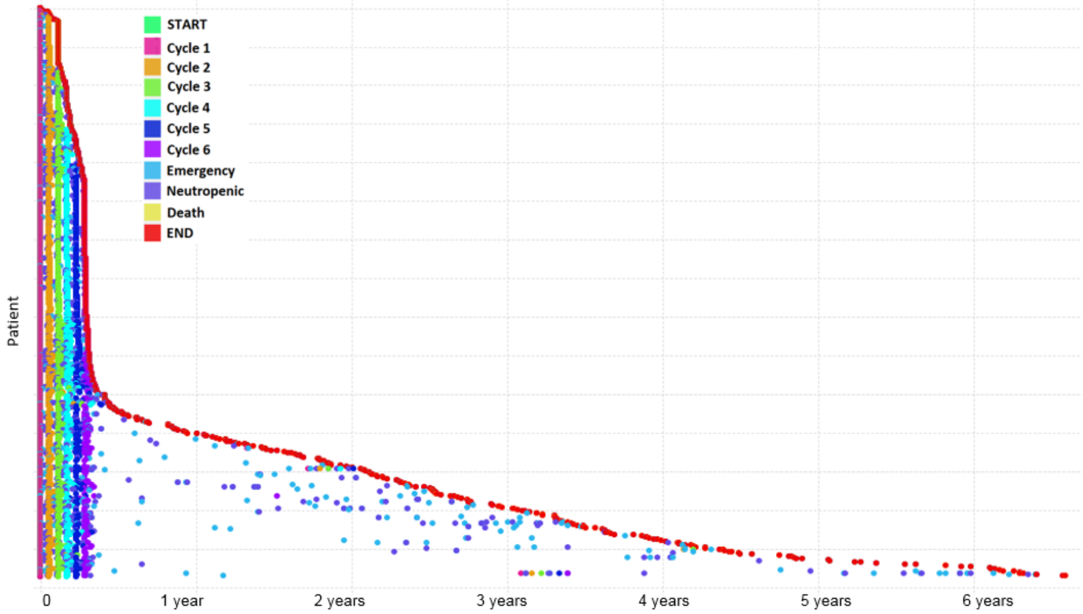


Figure 3.4 An example of a dotted chart (ProM). An activity is presented as a colour-coded dot, including the START and END activities. The x-axis shows the time relative to the start of the trace and the y-axis shows patient, ordered from the shortest to the longest duration.

A dotted chart can be created in ProM or bupaR. In both tools, an event log is needed as an input. In ProM 6.8, an event log can be visualised in a dotted chart by *Select visualisation > Dotted Chart (LogProjection)*. The resulted dotted chart can be adjusted by setting the *x-axis attribute, y-axis attribute, trace sorting, color attribute, shape attribute, attribute statistics,* and *connect/ disconnect events*. In bupaR, a dotted chart can be created using the following syntax:

```
event_log %>% dotted_chart(x=”absolute”, y=”start”)
```

4) Process comparison diagram

A process comparison diagram shows an annotated transition system resulting from the Process Comparator plugin in ProM, as presented in Section 2.2.6.1. Figure 3.5 shows a process comparison diagram from an experiment in Chapter 6.

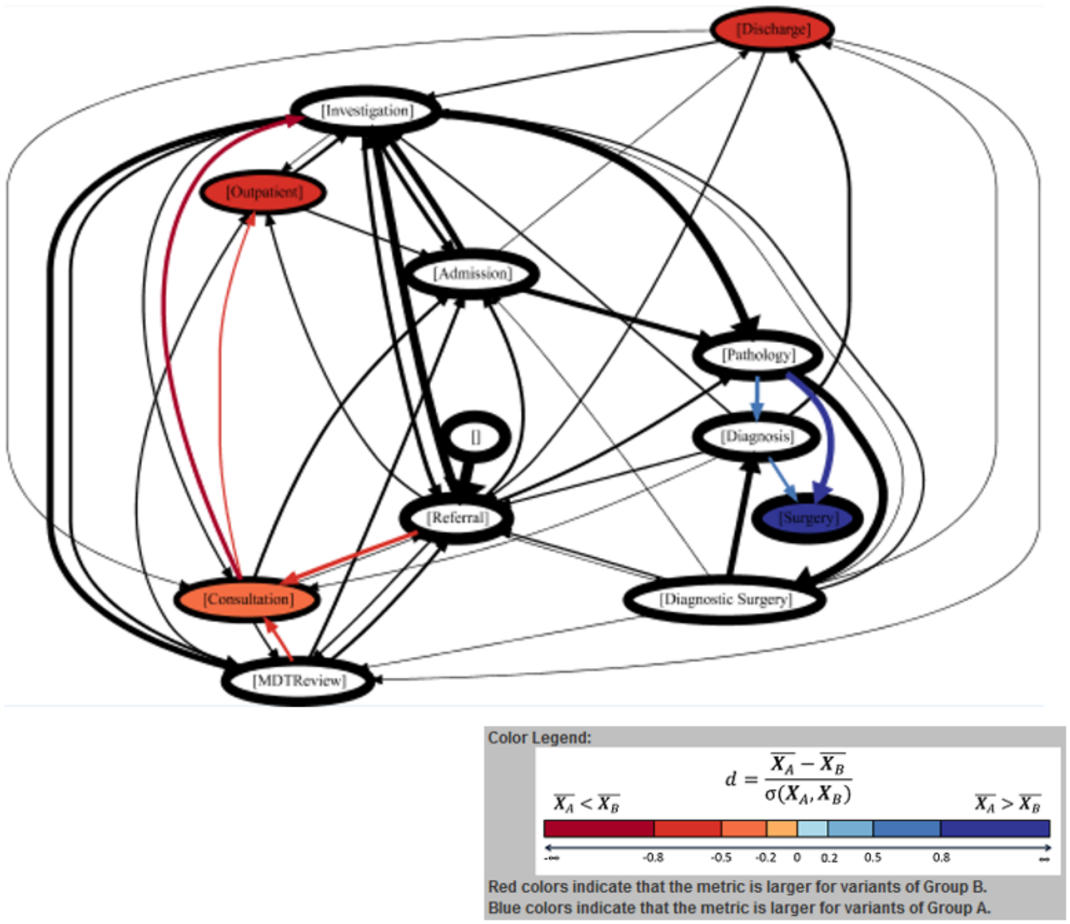


Figure 3.5 An example of a process comparison diagram. A node represents an activity and an arc represents path between activities. Node and arc thickness represent trace frequency. Colours represent the percentage differences in activities and paths in the two event logs. Blue means the percentage of trace frequency in the first group is higher than in the second group and red means the reverse.

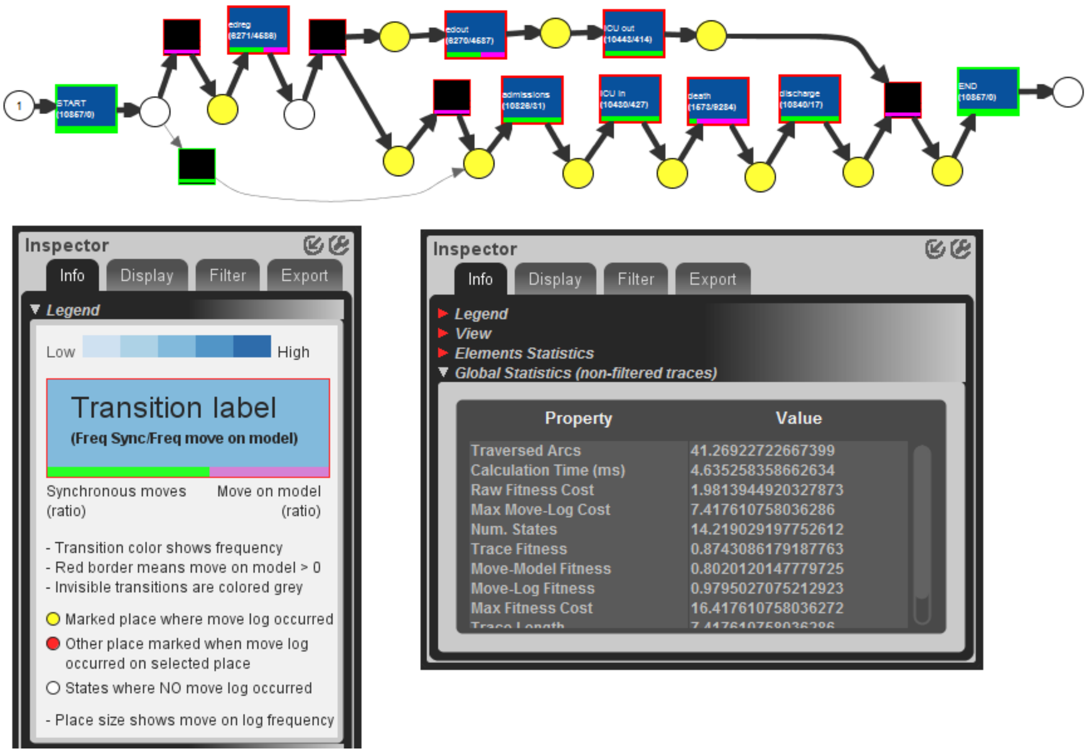
Figure 3.5 shows a process comparison diagram. In this study, the trace frequency is chosen. The alpha significance level was set as the default 5%. In ProM 6.8, this diagram can be built from at least two event logs using the *Process Comparator* plugin. The plugin asks to set Group A and Group B, then shows the graph. There are two settings that can be adjusted, which are transition system settings (graph properties and filter) and comparison settings (process metrics and alpha significance level).

3.2.3.2 Conformance checking

Conformance checking was done to check if the resulting model could represent the reality captured in the event log. Conformance checking can also be seen as an attempt to measure the quality of the process model resulted from the process discovery. The conformance checking was performed using some plugins in ProM, including:

1) ***Replay Log on Petri Net for Conformance Analysis***

This plugin accepts a Petri net and an event log. It provides conformance values based on cost-based fitness analysis [158, 159]. This plugin measures fitness that represents to what extent the process model captures the observed behaviour. This is achieved by identifying skipped and inserted activities. Skipped activities are activities that should be performed based on the model, but do not happen in the log. On the other hand, inserted activities are activities that occur in the log, but should not happen based on the model. Both skipped and inserted activities affect fitness value. The output is a Petri net annotated with replay results. For example, in Figure 3.6, Petri net can be presented with a legend and its global statistics.



Inspector

Info | Display | Filter | Export

▼ Legend

Low High

Transition label
(Freq Sync/Freq move on model)

Synchronous moves (ratio) Move on model (ratio)

- Transition color shows frequency
- Red border means move on model > 0
- Invisible transitions are colored grey

○ Marked place where move log occurred
● Other place marked when move log occurred on selected place
○ States where NO move log occurred

- Place size shows move on log frequency

Inspector

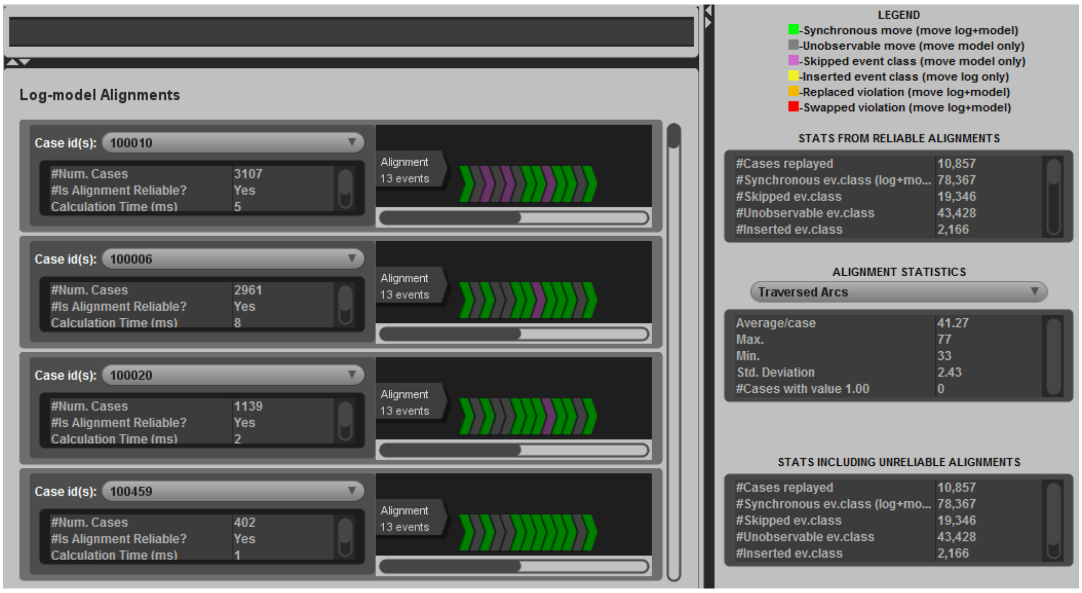
Info | Display | Filter | Export

► Legend
► View
► Elements Statistics
▼ Global Statistics (non-filtered traces)

Property	Value
Traversed Arcs	41.26922722667399
Calculation Time (ms)	4.635258358662634
Raw Fitness Cost	1.9813944920327873
Max Move-Log Cost	7.417610758036286
Num. States	14.219029197752612
Trace Fitness	0.8743086179187763
Move-Model Fitness	0.8020120147779725
Move-Log Fitness	0.9795027075212923
Max Fitness Cost	16.417610758036272
Trace Length	7.417610758036286

Figure 3.6 An example of a Petri net with a legend and its global statistics. The Petri net is annotated with replay results. The left window shows the log-model alignments, while the right window shows the list of deviations.

In Figure 3.7, the log-model alignment maps each case with the process model to show deviations. Some statistics are also presented, which include the statistics from reliable alignments and statistics including unreliable alignments. Other options of the visualising the results are 'time between transition analysis', 'trace alignment of alignments', 'visualise p-alignments as graphs, and 'visualise p-traces as graphs'.



Log-model Alignments

Case id(s): 100010

#Num. Cases	3107
#Is Alignment Reliable?	Yes
Calculation Time (ms)	5

Alignment 13 events

Case id(s): 100006

#Num. Cases	2961
#Is Alignment Reliable?	Yes
Calculation Time (ms)	8

Alignment 13 events

Case id(s): 100020

#Num. Cases	1139
#Is Alignment Reliable?	Yes
Calculation Time (ms)	2

Alignment 13 events

Case id(s): 100459

#Num. Cases	402
#Is Alignment Reliable?	Yes
Calculation Time (ms)	1

Alignment 13 events

LEGEND

🟩-Synchronous move (move log+model)
▨-Unobservable move (move model only)
▨-Skipped event class (move model only)
🟨-Inserted event class (move log only)
🟧-Replaced violation (move log+model)
🟥-Swapped violation (move log+model)

STATS FROM RELIABLE ALIGNMENTS

#Cases replayed	10,857
#Synchronous ev.class (log+mo...	78,367
#Skipped ev.class	19,346
#Unobservable ev.class	43,428
#Inserted ev.class	2,166

ALIGNMENT STATISTICS

Traversed Arcs

Average/case	41.27
Max.	77
Min.	33
Std. Deviation	2.43
#Cases with value 1.00	0

STATS INCLUDING UNRELIABLE ALIGNMENTS

#Cases replayed	10,857
#Synchronous ev.class (log+mo...	78,367
#Skipped ev.class	19,346
#Unobservable ev.class	43,428
#Inserted ev.class	2,166

Figure 3.7 An example of a replay result presented as the alignment to the log. The left window presents the log-model alignments. The list of deviations is in the legend in the right window.

The *'Replay a log on Petri Net for Conformance Analysis'* plugin can be used for checking conformance of a log to a process model in Petri net notation. The results can be presented in many ways, including the annotations in the Petri net along with detail information for each activity, each transition, and for the complete model. The information include calculation time, trace fitness, move-model and move-log fitness.

2) *Measure Precision/Generalisation*

This plugin provides precision and generalisation values as mentioned in [85]. This plugin measures the precision of a process model given an event log by first aligning the traces in the log to the model. In this study, the "Replay a log on Petri Net for Conformance Analysis" plugin was used by providing a Petri net, the Petri net replay result from point (1), and the event log. The results are presented simply as the precision and generalisation values.

For example, the event log from experiment 1 in case study 1 as discussed in Section 4.2 can be used as an input along with the Petri net resulted from the iDHM plugin as shown in Figure 3.6 and the Petri net replay results as presented in Figure 3.7. The results are a precision value of 0.86345 and a generalisation value of 0.99586.

The ***enhancement*** involved additional data, time, and performance perspectives to analyse the process model. An example of additional data perspective involved adding details from other tables in the database. Another example in the additional time perspective was the calculated time duration. In the performance perspective, the additional data was gathered from discussions with clinical experts to define several ways to describe performance of patient treatment.

The ***process analytics*** focused on process change analysis, following the approaches in concept drift analysis. More details about the process change analysis done in this research are presented in Section 3.3.

3.2.4 Evaluation

The objective of the evaluation stage was to ensure that the findings could answer all the original research questions and were meaningful for the domain experts. Evaluation of this study includes both statistical evaluation and clinical evaluation. The evaluation was effected through a series of focus group discussions in formal meetings. The statistical and clinical experts were invited to a discussion at the end of an iteration to verify and validate the results. Group discussions were also conducted during all the stages of the study to evaluate the approaches taken in each step and to gather inputs from the experts to enrich the next steps.

The statistical evaluation was performed to evaluate the results of the *mining and analysis* stage. Process models discovered in the process mining stage were evaluated using the conformance values of trace fitness, precision, and generalisation metrics using the plugins in ProM 6.8. Process change analysis was evaluated using t-test and hypothesis testing approaches. When a change point was detected, statistical evaluation was performed to test if the detected change was statistically significant. A complete statistical evaluation was done in the PPM Cancer experiments. It was performed based on the insights from the MIMIC-III and the PPM Chemotherapy experiments. Ciaran McInerney, Ph.D., a statistical analyst helped in the statistical evaluation in the PPM Cancer experiments.

The clinical evaluation was conducted through discussions with clinical experts. This was done in the PPM Chemotherapy and PPM Cancer analyses. Whenever required, other experts were also consulted. In the PPM Chemotherapy case study, there were also discussions with the previous research team, including Karl Baker and Elaine Dunwoodie. In the PPM Cancer case study, there were also discussions with the PPM training team, the PPM development team, and the Leeds Care Record (LCR) research team. The PPM development team members with whom discussions were held in this study included Colin Johnston, Nigel Stanworth (PPM Release Manager), and Jonny Smith. Discussions with the LCR research team was effected through Julia Millman, the program manager of LCR. The final evaluation was done by Professor Geoff Hall, a senior lecturer in Medical Oncology and Chief Clinical Information Officer at the LTHT.

3.3 Process change analysis

Process change analysis followed the approaches in concept drift analysis [16]. This step analysed the treatment processes and how they had changed over time. There are three main challenges in dealing with concept drift analysis. These are: change point detection, change localisation and characterisation, and change process discovery. The approaches taken in this study to address those challenges are as follows:

3.3.1 Change point detection

The first challenge was to detect a point in time where the process had been changed. A straightforward method to do this was to split the data based on a specific event as a reference. This was done in the analysis of the PPM Cancer data by partitioning the event log based on a pre-defined duration/ window size. In this study, the partitioning was done based on the diagnosis year, as illustrated in Figure 3.8.

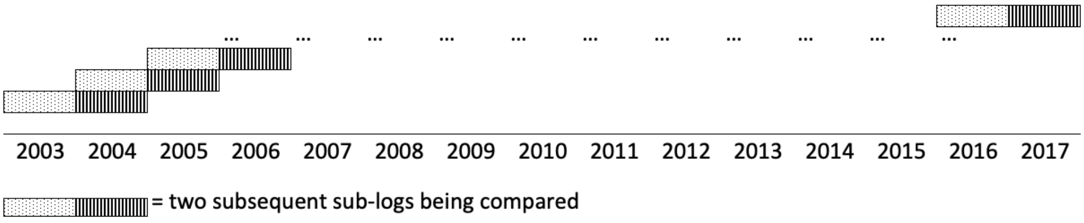


Figure 3.8 Illustration of log partitioning approach. Each year partition consists of traces of patients diagnosed in that particular year. For each iteration, two subsequent year partitions are being compared.

In addition to Definition 1 and 2 in Section 3.2.2, the following definition is applied:

> **Definition 3 (Log partitions)**. A partition ***P*** is a subset of an event log ***E*** based on partitioning criteria. The partitioning is done such that a trace is grouped into a partition with no duplication in other partitions. For this study, the partitioning was done based on the year of the timestamp ***t*** of the *diagnosis*. The event log was therefore split into partitions based on the year of diagnosis of each patient. There are clearly many partitioning options that could be adopted.

As presented in Section 2.2.5.2, some approaches have been proposed to detect change points. This study used a statistical testing approach to find significant differences between the processes. This approach is flexible enough to test differences based on various features describing processes.

In this study, change detection was performed in the PPM Chemotherapy and PPM Cancer analyses. Change detection was not possible in the MIMIC-III dataset due to the date shifting approach during the data creation by the MIMIC-III team. The documentation of the MIMIC-III data mentioned that the hospital information system from which the MIMIC-III data were collected had been changed and this was used as a starting point to analyse changes in the MIMIC-III database.

In the PPM Chemotherapy and PPM Cancer analyses, the PPM Electronic Health Record (EHR) from which the data was collected is continually growing. The development team works to make many changes in many aspects of the PPM EHR based on the changing needs of the clinical teams. In the PPM Chemotherapy analysis, change detection was done by comparing traces in different years. In the PPM Cancer analysis, change detection was done to a GP Tab change as a known change and by analysing monthly records to detect the unknown changes. Two different approaches in the PPM Cancer case study were the multi-level approach and a combination approach of signal decomposition and Statistical Process Change (SPC) chart. Each of them will be described in the following sections.

3.3.1.1 A multi-level approach for identifying process change

In this approach, the process comparison was done using several metrics at three different levels: process model level, trace level, and activity level. For the model-level comparison, a general process model was built using interactive Data-Aware

Heuristics Miner (iDHM) in ProM from the complete event log. iDHM was chosen because this plugin supported interactive adjustment to the level of details needed in the analysis. The model-level behaviour was described by the conformance values in the replay fitness, precision, and generalisation of each sub-log to the general model. The trace-level behaviour was described by durations and the proportion of trace variants in the sub-logs. The activity-level behaviour was described by activity frequency and its percentage in the sub-logs.

The metrics for those three levels of process comparison are presented in Table 3.2. This approach was used in a case study of endometrial cancer pathways from GP referral to the diagnosis of cancer, as presented in experiment 6 of the PPM Cancer case study in Section 6.3.

Table 3.2 Metrics for multi-level process comparison

Level	Metrics	Description
Model	Replay fitness	The ability of the model to accurately reproduce the traces recorded in the log.
	Precision	The proportion of the behaviour allowed by the model which is not seen in the event log.
	Generalisation	The ability of the model to reproduce the future behaviour of the process.
Trace	Duration	The number of days of the pathway from Referral to Diagnosis.
	Variant proportion	The proportion of variants in the sub-log that were one of the most frequent variants in the complete log.
Activity	Frequency	The number of patients having a specific event within one year.
	Percentage	The percentage of patients having a specific event out of all patients within a year.

3.3.1.2 Signal decomposition and Statistical Process Control (SPC) chart for identifying process change

This approach identifies process change based on the pattern of monthly records in the EHR system. The hypothesis is that it is possible to detect change points based on the monthly records over time. A plot was created based on the observed number of monthly records per activity. The observed plots were decomposed using a signal decomposition technique [160]. The underlying assumption of this approach is that the observed plot can be decomposed to find the trend, seasonal, and remainder patterns over time, such that: $y_t = T_t + S_t + R_t$, where y_t is the observed data, T_t is the trend component, S_t is the seasonal component, and R_t is the remainder/ random component, all at period t. The remainder plot was then being analysed using the Statistical Process Control (SPC) chart [161] to get the change points with statistically significant different values from the previous period.

Two main functions in R were used in this approach: (1) the *decompose()* function in the fpp2 package for signal decomposition, and (2) the *qic()* function in the qicharts2 package for SPC chart. The decompose() function accepted a time series object of the observed data. For example, the monthly frequency of an activity can be defined and plotted as follow.

```
# define and plot time series data
ts_data = ts(data$monthly, start=c(year_start,month_start),
end=c(year_end,month_end), frequency=12)
```

The signal decomposition processed the monthly records of an activity to separate the trend, seasonal, and remainder patterns in the observed data. This study uses the classical additive decomposition to separate trend and seasonal patterns from the remaining/random data.

The ***trend*** pattern was based on the moving average smoothing, i.e. m-MA, meaning a moving average of order *m*. An *m*-MA can be written as $\hat{T}_t = \frac{1}{m}\sum_{j=-k}^{k} y_{t+j}$, where *m = 2k+1*. The estimate of the trend-cycle at time *t* is based on the average values of the time series within *k* period of *t*. The idea is that the observations during the nearby periods are likely to be close in value. The average is therefore eliminating some of the randomness in the data. A trend can be a long-term increase or decrease in the data, or a 'changing direction' when it goes from an increasing trend to a decreasing trend or vice versa. In this study, the trend pattern was based on the 12-MA smoothing, meaning that an average was counted from 6 months before and 6 months after a specific month. The consequence of this was that no value was calculated in the first 6 months and the last 6 months of the duration of the study. Other options to use 3-MA, 4-MA, and 6-MA were also tested but did not improve the results. This is why the 12-MA was chosen, which also implies the yearly average of monthly frequency.

The ***seasonal*** pattern is a fixed and known frequency. In this study, seasonal pattern is the monthly frequency. It was chosen to make sure that enough data are available in each season (month) while the individual data is not de-identified. Other options to analyse seasonal pattern in the daily, weekly, or quarterly frequency are also possible.

The observed data were de-trended by subtracting the trend component from the observed data. The average was then calculated from the same period over the years, such that $\hat{S}_t = y_t - \hat{T}_t$. For example, the seasonal component for January is the average of all de-trended January values over the years.

The ***random/ remainder*** component was calculated by subtracting the estimated seasonal and trend-cycle components: $\hat{R}_t = y_t - \hat{T}_t - \hat{S}_t$. It represents the observed signal after subtracted by the trend and seasonal patterns. For example, signal decomposition of a monthly frequency of an activity is as follows:

```
#signal decomposition
decompose_data = decompose(ts_data, "additive")
plot(decompose_data)
```

An example result is shown in Figure 3.9.

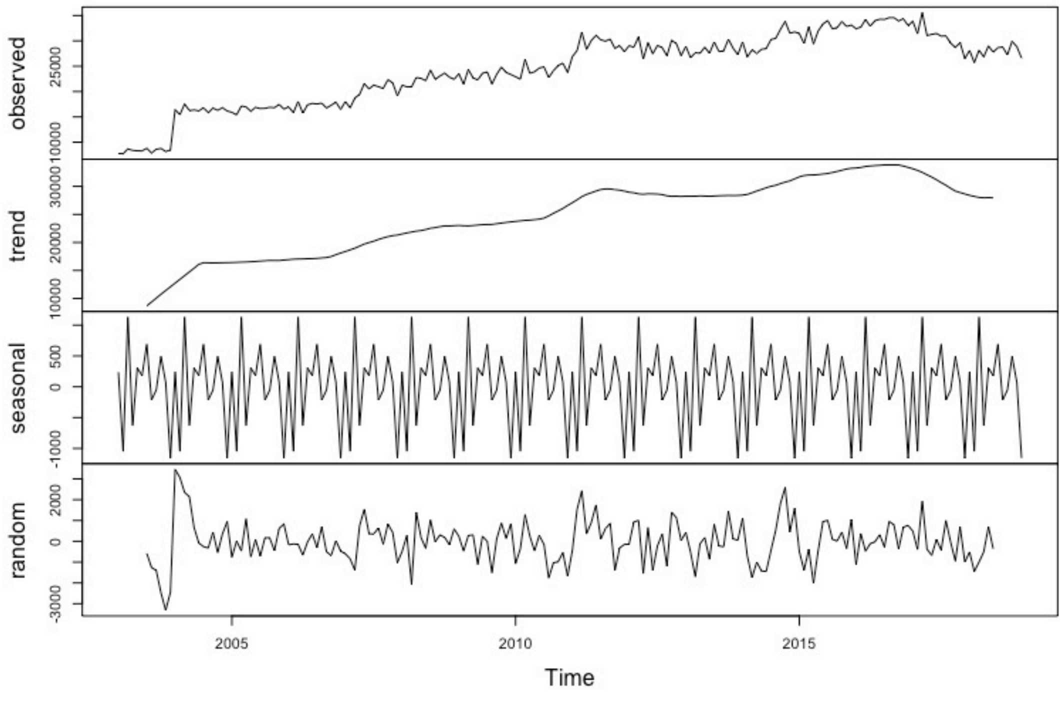


Figure 3.9 An example of a signal decomposition result. It shows the observed pattern, the trend pattern, the seasonal pattern, and the random/ remainder pattern. The x-axis is time and the y-axis is variance over the means.

The random/ remainder component was further analysed using ***the Statistical Process Control (SPC) chart***. SPC is commonly used for understanding process variation over time [156]. The remainder pattern was plotted to see the variability of the monthly records. The signals were compared to the upper and lower control lines. Change points were detected where the signal varied outside the control lines. This approach was used in an experiment on cancer pathways, as presented in experiment 8 of the PPM Cancer case study in Section 6.5.

In this study, the SPC chart was built using *qicharts2* library in R. The main function used in this study is *qic()* function, which creates run charts from time series data. A run chart is a point-and-line graph showing measures or count over time. Three horizontal lines are presented as the control lines. A centre line (CL) expressing the median, while upper control limit (UCL) and lower control limit (LCL) represent the upper and lower boundaries of variations in the data. The control limits are placed at ± 3 standard deviations from the centre line. Variation of values between LCL and UCL is considered as common cause variation that is present in any process, which is also called random variation or noise. Values outside the the control lines are considered as non-random variation or signal, which is caused by phenomena that are not normally present in the system. For example, the SPC chart of the remainder pattern of a monthly frequency of an activity can be resulted from a code as follow:

```
#input:
data_df <- ts_reshape(dec_data$random, type="long")
qic(value, data=data_df, chart="i", ylab='Count', xlab='Month')
```

This syntax resulted in an SPC chart as presented in Figure 3.10.

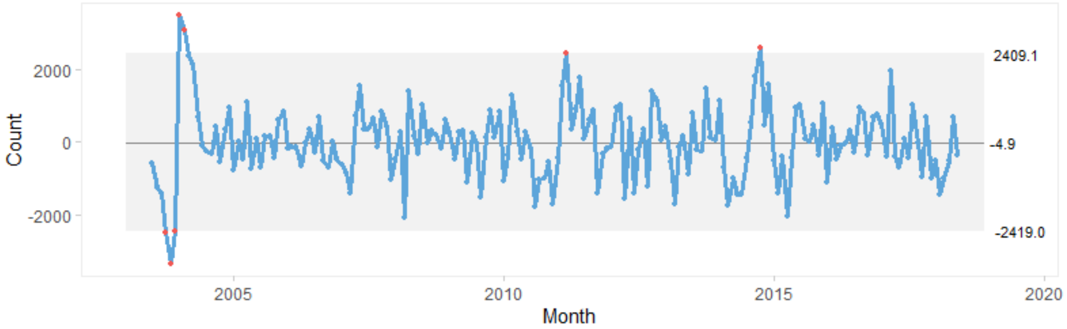


Figure 3.10 An example of the SPC chart. It shows the variability over the means of the random/ remainder pattern. The x-axis shows month and the y-axis is count. Red dots represents the points where there were significant changes in the variability over means.

3.3.2 Change localisation and characterisation

After a change point was detected, the next task was to localise and characterise the change. This task involved both the identification of change perspective (for example: control-flow, data, resource, sudden, gradual, recurrent, incremental) and the exact change itself. Different types of data and process change require different techniques in change localisation and characterisation. In this study, the method for change localisation and characterisation was built and improved in the three datasets.

Change localisation and characterisation stage in the MIMIC-III data consisted of model-based and log-based comparisons. The model-based comparison was done using the DifferenceGraph plugin in ProM. This plugin compares two process models and visualises the differences between those two models. The log-based comparison was done using the Process Comparator plugin in ProM. This plugin compares trace frequencies in two event logs and visualises the significant differences. This stage in the PPM Chemotherapy data was done following the conformance-based change detection. Sub-logs created based on the diagnosis years were then being compared to each other on their trace fitness, precision, and generalisation. This stage in the PPM Cancer data was done based on the metrics for multi-level process comparison, as presented in Table 3.2. For change point detection using the multi-level approach, this stage was required to be done for each metric in the three different levels. A specific challenge in this task was to identify a significant change in the process based on those metrics in the three different levels of process comparison.

3.3.3 Change process discovery

Having now identified, localised, and characterised the changes, the next step is to relate the findings with the discovery of the change process to unravel the evolution of the process. This step can be done by analysing different perspectives of the changing process, such as using the performance metrics. The most important method in this step is the frequent focus group discussions with the domain experts.

The detected change points were discussed with the clinical experts. The discussions focused on the possible nature of the changes. The changes can be caused by several reasons, such as a change in the organisational structure in the hospital, a change in the guideline for cancer treatment, a change in the technology used within the hospital information system, or a change in the people who did the treatment process.

In this study, the change process discovery was done in the PPM Chemotherapy and PPM Cancer case studies. The change process discovery was conducted through discussions with clinical experts and the development team. Change process discovery was not possible in the MIMIC-III analysis because there is no direct access to clinical experts and because of the date shifting approach for anonymisation.

3.4 Case studies

Three case studies were built from the three datasets in this study: the MIMIC-III, the PPM Chemotherapy, and the PPM Cancer datasets. Each one of these is introduced in this section and explored in more detail in Chapters 4 to 6.

3.4.1 Case study 1: Experiments using the MIMIC-III dataset

The MIMIC-III dataset is a large, single-centred database comprising information relating to patients admitted to the Intensive Care Units (ICUs) in the Beth Israel Deaconess Medical Center (BIDMC) [25, 26, 168]. The MIMIC-III database was developed and has been maintained by the Laboratory of Computational Physiology at Massachusetts Institute of Technology (MIT) since 2003 [25, 26]. This database has been de-identified by removal of all protected health information. Public access to this data was made available through the National Institutes of Health (NIH).

The data covers 53,423 distinct hospital admissions for 38,597 adult patients (aged 16 years or above) admitted to the ICUs between 2001 and 2012. It contains data such as medication, laboratory measurements, charted observations during a patient's stay in the intensive care unit, and de-identified notes of the patient's stay. One important challenge was to select suitable tables for process mining. This challenge and how it was addressed in this study are described in Chapter 4. There are 16 out of 26 tables in the MIMIC-III database containing timestamped events. Those tables were used to discover process models based on the specific cohort of interest. The other 10 tables were used as reference tables during the analysis, such as PATIENT table to support analysis of the selected patients in a cohort.

Two different systems were in place over the data collection period. The Philips CareVue Clinical Information System (CV) was used between 2001–2008 and the iMDsoft MetaVision ICU (MV) was used between 2008–2012. This condition leads to the opportunity to compare the process models of those two systems. The MIMIC-III team provided the database combining data from both systems. In this study, clinical data from two systems were separated to be compared one another and analysed further with process mining approach. One limitation of using this dataset is that the deidentification process obscured the real dates. Another limitation is that there was no direct access to the clinical experts of this hospital, so that it was not possible to discuss the findings with the clinical experts. Despite the limitations, the

MIMIC-III dataset was useful to build initial method for process change analysis using process mining in this study, because (1) the MIMIC-III dataset is publicly accessible, therefore supports reproducibility of related research using the same dataset, and (2) there was a known system change to replace the CV system with the MV system, so that change characterisation and localisation can be done.

3.4.2 Case study 2: Experiment using the PPM Chemotherapy dataset

The Patient Pathway Manager (PPM) Chemotherapy dataset is the clinical data of cancer patients receiving chemotherapy treatment in the Leeds Cancer Centre during the period 1996–2015. This dataset was used in two previous studies in the research group entitled (1) "Using Routine Clinical Dataset to Develop Risk Algorithms in Oncology" (Ref: 13/NS/0128) [34] and (2) "Profiling Neutrophil Counts in Patients with Cancer During Cycle One of Chemotherapy" (IRAS ID 207804). The PPM Chemotherapy dataset consists of the clinical data of 31,511 patients in 13 tables. It was de-identified for the purpose of the previous study. There is no timestamp identified, but there is the possibility to obtain the sequence based on *Age* (in the number of days) and *Years*, which enable a historical pathway to be extracted for every patient. Those tables are related with a Patient ID (PID) which identify patient IDs. Later during this study, the dataset was backed up in a secured external driver for easier and safer access by restricted researchers.

The data came through automated extraction processes from patient hospital records and financial data held at the Leeds Teaching Hospitals NHS Trust (LTHT) to support patient care. Data analysts within LTHT provided non-identifiable data from records of all cancer patients in the PPM EHR system used by the hospital. The non-identifiable data had been encrypted and transferred to a secure environment – the University of Leeds Integrated Research Campus (UoL IRC). There was no known change identified at the beginning of the experiment of this case study.

In this study, the PPM Chemotherapy data was used to analyse six cycles of adjuvant chemotherapy in patients diagnosed with breast cancer between 2004 and 2013. Process mining was used to highlight variations from standard pathways of chemotherapy including the evidence of incomplete treatment and adverse events. This has also shown changes in the pathway over time.

The limitation of using these data in this study is related to the anonymisation undertaken during data collection, where all dates were replaced with the patient age as a number of days. This anonymisation approach makes it impossible to include the fine-grained level of dates. For example, it was not possible to analyse busy days in the hospital, because there was no real date included in the anonymised dataset. Despite this limitation, the number of days recorded for each event can be used to infer sequence of events in patient treatments.

3.4.3 Case study 3: Experiments using the PPM Cancer datasets

The Patient Pathway Manager (PPM) is the EHR system developed and used in the LTHT. It holds the records of more than 3 million patients. Initially developed in 2003 [165] to support the collection of key information for the National Cancer Dataset and reporting of Cancer Outcomes Services Dataset [166]. The PPM system was built around a standard SQL Server 2005 infrastructure and is viewed and administered via a User Interface (UI) application constructed and developed using Visual Basic. It was then extended by LTHT in 2010 as a web-based Electronic Patient Record (EPR) and known as PPM+. The web-based PPM Portal went live in March 2012. This system integrates electronic data held within the Trust into a single EHR database.

PPM+ is the current development of the PPM system which delivers the EPR for LTHT staff and the Leeds Care Records (LCR). LCR integrates patient records across health and social care organisations in Leeds citywide. The PPM system was developed into PPM+ and used within Trust since 2013. The PPM+ accesses the same database of the PPM with additional connections to other IT systems, including primary care data. Following common terms used in the LTHT, the web-based PPM+ is now referred to as PPM and the older version of PPM is referred to as PPM1.

In this study, the data is extracted from the PPM Query database, a copy of a real-life PPM database. The ethics approval of the PPM Cancer dataset was through an NHS honorary contract and an Integrated Research Application System (IRAS) to gain the Health Research Authority (HRA) Approval (REC Reference 18/HRA/0410). LTHT provided non-identifiable data from records of a group of cancer patients. An automated process was used to produce the non-identifiable row-level data in line with the Information Commissioner’s Office (ICO) and NHS standards.

Along with the data from the PPM Query database, this study used data from the PPM Splunk and PPM JIRA. The PPM Splunk records all user access to the PPM system, which is useful in analysing system usage for specific functionalities being examined. The PPM JIRA contains details of changes that had happened to the system, including the type of changes, contents of the release when the change(s) were applied, when they were applied, and any supporting training notes.

In this study, the PPM Cancer dataset was used to analyse process change over time. One experiment explored a process before and after a known change happened, while two experiments were done to explore process change without a known change at the beginning of the experiment. The limitation of using this dataset is that the growing nature of the database makes it difficult to analyse the data alone without analysing documentations of the PPM system and discussing with clinical experts and the development team. The documentation and discussions were needed to reveal the changes that happened to the system during the long period of data collection.

3.5 Summary

This chapter has explained the general methodology and datasets used in this study. The main stages in the general methodology were built based on the L* life-cycle model and the PM^2 method. The four main stages were: (1) planning and justification, (2) ETL, (3) mining and analysis, and (4) evaluation. Additional methods were followed to complete some other steps within the stages, which are the question-based methodology, the ClearPath method, the concept drift analysis, and signal decomposition and SPC methods. The ClearPath method has been published [91] and cited in 11 articles. Some of the steps in the general methodology were not applicable in the case studies because of the specific limitations of each case study.

The next three chapters present the three case studies based on the three datasets used in this research. All three case studies apply the methodologies that has been presented in this chapter. Chapter 4 describes the MIMIC-III data as the first case study. Chapter 5 describes the PPM Chemotherapy data as the second case study. Chapter 6 describes the PPM Cancer data as the third case study.

Chapter 4
Case study 1: Experiments using the MIMIC-III dataset

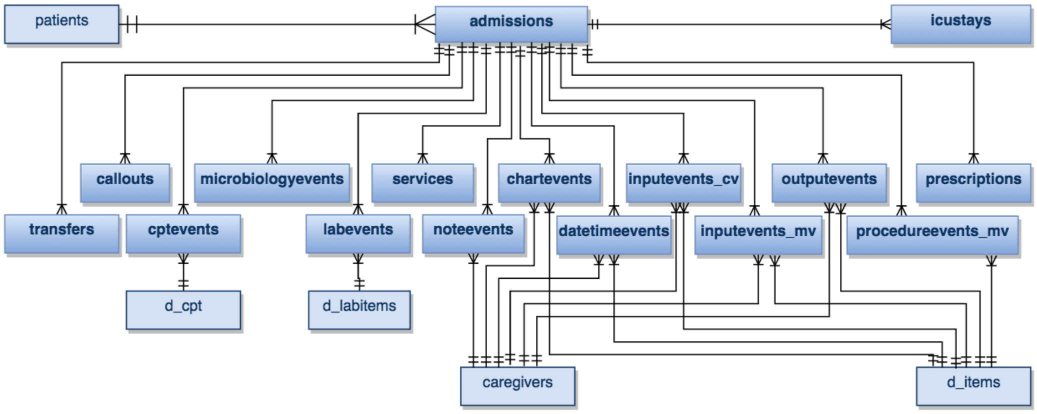
The MIMIC-III is the first case study and is presented in this chapter. The data quality assessment part in Section 4.1.3 has been presented in a poster at the Informatics for Health 2017 conference in Manchester, UK. The abstract was published in the Journal of Innovation in Health Informatics [167]. The work was extended and published in a full journal paper entitled "The assessment of data quality issues for process mining in healthcare using MIMIC-III, a freely available e-health record database" in the Health Informatics Journal [168]. Section 4.2 has also been presented in an IEE conference in Indonesia and published in a conference paper entitled "Process mining in oncology using the MIMIC-III dataset" in the Journal of Physics: Conference Series [169]. Part of the work in this chapter was also presented in a joint presentation in the workshop of the Worldwide Universities Network (WUN) 2017 in New York, USA.

4.1 Data description

Overview of the MIMIC-III dataset has been presented in Section 1.4.2 and has been described in Section 3.4.1. More details of this dataset are presented in this section.

4.1.1 Data characterisation

The data source of the MIMIC-III dataset has been described in Section 1.4.1, and the overview of the dataset has been described in Section 3.4.1. The concept-level Entity-Relationship (E-R) diagram of the MIMIC-III database is displayed in Figure 4.1.



event-related entities are in* **bold

Figure 4.1 The concept-level E-R diagram of the MIMIC-III database. Five reference tables are: drgcodes, d_icd_diagnosis, d_icd_procedures, diagnosis_icd, procedure_icd.

Figure 4.1 shows the concept-level E-R diagram of the MIMIC-III database. Time in the MIMIC-III database is stored with one of two suffixes: TIME (down to the minute) and DATE (down to the day). There is also *charttime* indicating when the observation was made and *storetime* indicating when it was validated. In this study, the event logs were created using *charttime* attributes, as this is closest to the time of actual measurement. All patient data in the MIMIC-III database were de-identified and all dates randomly shifted to the future so that dates are internally consistent for the same patient but inconsistent across patients.

4.1.2 Scope

In this study, the MIMIC-III v1.3 was used. This version was released on 10 December, 2015 [162]. It contains a wide range of data such as the admissions and discharges, ICU stays, laboratory measurements, outpatients, and charted observations during patient stays in the ICU. Data were curated by the MIMIC-III team from archives of critical care information systems, hospital Electronic Health Record (EHR) databases, and the Social Security Administration Death Master File. The rich nature of the MIMIC-III dataset provides the timestamped data of clinical events in the 16 event tables, which is suitable for process mining.

In this study, the MIMIC-III database was used to provide a test case and to build a methodology for healthcare process mining. The patients from the MIMIC-III database were included if they were diagnosed with colorectal cancer, at least once. Colorectal cancer was chosen as a case study because this is one of the most common types of cancer. Colorectal cancer has already been discussed in Section 2.1.6.1 by providing related pieces of literature.

4.1.3 Data quality

The data quality assessment of the MIMIC-III dataset was done following the Weiskopf & Weng framework [133]. The data quality was assessed with a specific focus for process mining projects. The data quality assessment, as presented in the journal paper, was conducted following the L* life-cycle model [3] with an adaptation before the extraction stage. This was done by reconstructing the database in a local database management system (PostgreSQL). The idea was to get the fullest possible dataset from the MIMIC-III database that can be iteratively extracted later to create smaller subsets, mined, and assessed for the quality.

Five out of the seven methods suggested in the Weiskopf & Weng framework have been used to assess the data quality of the MIMIC-III database for process mining. Those methods resulted in some findings in four out of five dimensions of data quality. Two methods (gold standard and log review) were not applicable due to the anonymisation process done by the MIMIC-III database provider. A summary of the data quality assessment is presented in Table 4.1.

Table 4.1 Data quality assessment of the MIMIC-III database

dimension method	completeness	correctness	concordance	plausibility	currency
Element presence	*Y*	*Y*	*Y*	*Y*	*N*
Data element agreement	*Y*	*N*	*N*	*N*	*N*
Data source agreement	*Y*	*Y*	*Y*	*N*	*N*
Distribution comparison	*Y*	*N*	*Y*	*Y*	*N*
Validity check	*Y*	*Y*	*N*	*N*	*N*
Gold standard	<<*not applicable*>>				
Log review	<<*not applicable*>>				

**Y = yes (applied) *N = no (the method cannot be used to assess the dimension)*

4.1.3.1 Element presence

Element presence was done by checking the presence of the three minimum attributes for process mining: *case_id*, activity and timestamp. The *case_ids* are available in the 16 event tables are the *subject_id* and *hadm_id*. Another possible case id is *icustay_id,* but it is not available in seven tables (*admissions, callout, cptevents, labevents, microbiologyevents, noteevents,* and *services*). One patient (*subject_id*) might have more than one hospital admission (*hadm_id*) and one admission might have more than one ICU stay (*icustay_id*). The *icustay_id* is only available for events recorded during ICU stays. Those ids represent the event granularity options for process mining that process miners should be aware of in the analysis.

The *activities* are available in all 16 event tables. Activity names are recorded directly in nine tables (*admissions, callout, cptevents, icustays, noteevents, prescriptions, microbiologyevents, services,* and *transfers*). The activity names were referred from the *d_items* table in the other six tables (*chartevents, datetimeevents, inputevents_cv, inputevents_mv, outputevents,* and *procedureevents_mv*). There are two levels of granularity in the *d_items* table: label (fine-grained level) and category (coarse-grained level). Another table (*labevents*) has to refer to the *d_labitems* table.

The *timestamps* are available in all 16 event tables. The date versus time and *charttime* versus *storetime* issues, as presented in Section 4.1.1, are important in process mining projects. The different granularities of the timestamps (*date* – down to the day, and *time* – down to the minute) presents some issues in process mining. When tables having different granularity of timestamps are combined, the sequence of the events and the process duration might be incorrect. Four tables can be used to analyse activity duration because they recorded the start and end times. Those tables are the *icustays, inputevents_mv, procedureevents_mv,* and *transfers* tables.

4.1.3.2 Data element agreement

Data element agreement was done to compare two or more elements in the database to see if they report the same or have compatible information. In this study, the data element agreement was done by tracing back to the MIMIC-III website and the data descriptor. The preliminary assumption was that the data descriptor described the MIMIC-III database accurately, but this was not always the case.

There are three findings related to completeness of the *case_ids*, level of detail of the timestamps, and plausibility of the data. Completeness of the case ids was found to be dependent on the *case_id* chosen for the analysis. When the *case_id* is *subject_id* or *hadm_id*, the *admissions* and *transfers* tables are complete. But when the case id is *icustay_id*, the *icustays* and *transfers* tables are complete. Depending on the *case_id* used in the analysis, completeness of all other tables can be checked by reference to those tables. The plausibility issues have been found by comparing the data duration with the MIMIC-III data descriptor. The MIMIC-III data descriptor specified that the dates had been shifted into the future to the years 2100 to 2200. It was found that some events were dated before 2100 and after 2200. This might be caused by historical data such as test results and scheduled events, such as scheduled treatments. In this study, there are no immediate data quality issues, but this could cause an error if the events were selected by a date range.

4.1.3.3 Distribution comparison

Distribution comparison was done to compare the records in the MIMIC-III database to the data descriptor [162]. This was done to check completeness, concordance, and plausibility of the event tables in the MIMIC-III database.

The *subject_id* is complete in all event tables, but the *hadm_id* and *icustay_id* are not. There are 70 missing *spec_itemid* in the *microbiologyevents* table, but those can be replaced entirely by *spec_type_desc*. In the admissions table, there are missing timestamps (37%) that represent patients who are not dead or not admitted in the Emergency Department (ED). Without access to the data source, there is no way to decide if the missing values are because the events did not happen or because they were not recorded correctly.

Some tables have missing timestamps: *callout* (26%), *cptevents* (82%), *microbiologyevents* (7.5%), *noteevents* (15%), and *transfers* (11%). The MIMIC-III documentation mentions that the collection of *callout* data only began part way through the MIMIC-III database and with date shifting this missing data has been spread at random. Incomplete timestamps in the *charttime* of *microbiologyevents* and *noteevents* tables can be derived by linking to *chartdate* but consider that the granularity level would be different. This incompleteness also happened in *cptevents, prescriptions*, and *transfers* tables so process mining would be unreliable.

4.1.3.4 Validity checking

Validity checking was done by querying data from each table and between related tables, to determine if the values ‘make sense’ in the problem domain. The findings were related to the ICD-9 codes and duplicate records between different tables.

The MIMIC-III database provides reference tables, which are *d_icd_diagnosis* describing the diagnosis codes and *d_icd_procedures* describing the procedure codes. The analysis of those two tables found that 144 out of 14,711 (0.98%) diagnosis codes are missing and 16 out of 258,082 (0.01%) procedure codes are missing. The percentages of the missing codes are small, but they can be significantly affected the analysis of a specific cohort containing those missing codes.

There was also a duplication problem between different tables, for example, *datetimeevents* and *admissions* tables. In the *datetimeevents* table, there is a ‘hospital admit date’ that is duplicated with *admittime* in the *admissions* table. For those duplicates, 1,696 out of 24,549 (7%) admissions are matched, while 22,658 out of 24,549 (92%) have earlier admission dates in the *admissions* table compared to those in the *datetimeevents* table. In this case, the admission dates in the *admissions* table was selected, and the duplicated records in the *datetimeevents* table were ignored.

4.1.3.5 Data source agreement

Data source agreement was checked to compare data from two sources of the MIMIC-III database, which are CV and MV. The CV system was used during 2001–2008 and the MV system was used during 2008–2012. The CV system was provided by Philips and was used to archive clinical data at the bedside of the ICU patients admitted to the BIDMC during 2001–2008. The MV system was used to archive the clinical data of patients admitted during 2008–2012. It is important to note that the duration of the records in the CV system (8 years) was longer than the duration of the records in the MV system (4 years). Patients in those two systems have data archived in different formats. Another consequence is that the patients recorded in the CV system have records from a longer duration than those in the MV system.

In this study, it is important to know that the data source of the MIMIC-III database had been changed. A backward approach has been used to create separate event logs from the two EHRs. Those event logs were then used to discover two process models that could then be compared. This is described in more detail in Section 4.3 (Comparing CV and MV systems).

4.1.4 Representativeness

Patients were included in the MIMIC-III database if they had at least one ICU stay. All clinical data for those patients were also included. This database is therefore a comprehensive example of EHR data from a large hospital. Representativeness of this database has been explored based on age group at first admission.

A description of the MIMIC-III dataset was presented in Section 3.4.1. Figure 4.2 shows the age groups at first admission. There were 1,991 patients (4.28%) excluded because their age was calculated to be over 300 at their first admission. This limitation is due to the time-shifting done by the MIMIC-III team to protect patient confidentiality. The MIMIC-III data description also mentioned that patients who were older than 89 years at any time in the database have had their date of birth shifted to obscure their age.

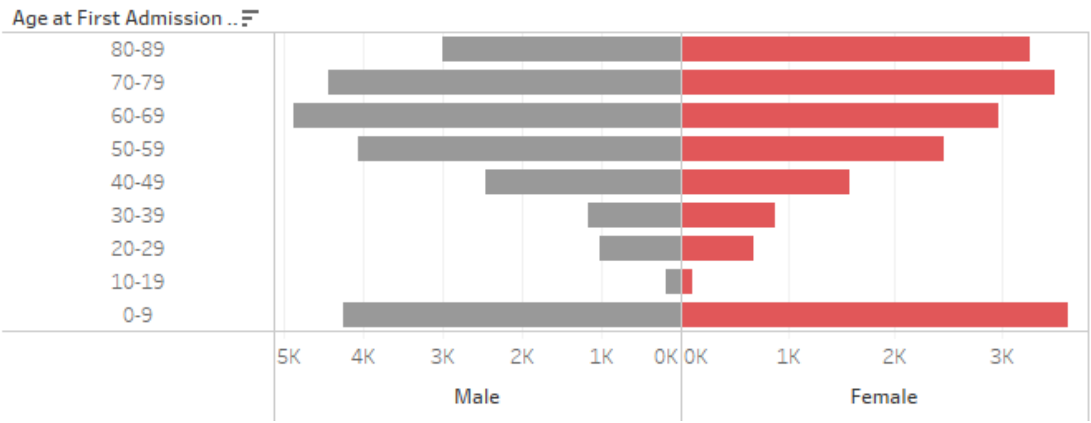


Figure 4.2 Distribution of age groups at the first admission. The x-axis shows the number of patients (in thousand/ K) and the y-axis shows the age groups at first admission. Gender are colour-coded with Male on the left and Female on the right.

Figure 4.2 shows the distribution of age groups at first admission. Of a total of 46,520 patients in the MIMIC-III dataset, there were 7,872 neonates (18%) and 36,674 adult patients (82%) aged 16–89. There were 25,424 male (57%) and 19,150 female (43%) patients. Those numbers have been checked to be in agreement with the MIMIC-III data descriptor. The patients in the 0–9 years group are from the CV system, which included newborns.

Representativeness of the MIMIC-III dataset is high. Figure 4.2 shows that all patient groups and both genders are represented in the data. There was a small number of patients in the young age group (10–19), which corresponds to the real-life situation. However, a result of this small number of young patients means that the analysis of this cohort may not be robust.

4.1.5 Data variety

The data variety of the MIMIC-III dataset can be explored by describing the variety of diagnosis and procedure codes of the patients. The distribution of patients in 18 diagnosis groups is presented in Table 4.2. It shows that all groups of diagnoses are represented in the MIMIC-III patients, allowing a range of analysis from ICD code 001 to 999, and E&V. Each patient had 0 to 144 different diagnosis codes. The most common diagnosis group is the *External causes of injury and supplemental classification* (n = 34,457/ 74.1%), followed by the *Diseases of the circulatory system* (n = 32,503/ 69.8%) and the *Endocrine, nutritional and metabolic disease, and immunity disorders* (n = 27,440/ 58.9%). There are 47 patients had a null (0.10%).

Table 4.2 Distribution of diagnosis groups. A patient might have more than one diagnosis and included in more than one group. Percentage (%) is calculated over the total patients.

ICD code	Diagnosis Label	n	%
001-139	Infectious and parasitic diseases	11,577	24.89
140-239	Neoplasms	7,361	15.82
240-279	Endocrine, nutritional and metabolic diseases, and immunity disorders	27,440	58.99
280-289	Diseases of the blood and blood-forming organs	15,661	33.64
290-319	Mental disorders	13,400	28.80
320-389	Diseases of the nervous system and sense organs	12,744	27.39
390-459	Diseases of the circulatory system	32,503	69.87
460-519	Diseases of the respiratory system	19,973	42.93
520-579	Diseases of the digestive system	16,730	35.96
580-629	Diseases of the genitourinary system	16,765	36.04
630-679	Complications of pregnancy, childbirth, and the puerperium	161	0.35
680-709	Diseases of the skin and subcutaneous tissue	5,097	10.96
710-739	Diseases of the musculoskeletal system, connective tissue	8,391	17.97
740-759	Congenital anomalies	2,990	6.43
760-779	Certain conditions originating in the perinatal period	5,321	11.44
780-799	Symptoms, signs, and ill-defined conditions	16,910	36.35
800-999	Injury and poisoning	19,318	41.53
E&V	External causes of injury and supplemental classification	34,457	74.07
Null		47	0.10

The distribution of patients within the 18 procedure groups is presented in Table 4.3. It shows that all groups of procedures are represented in MIMIC-III, allowing the analysis of many types of procedures, from procedure 01 to procedure 99. Each patient had 1 to 98 different procedure codes. The most common procedure group is the *Miscellaneous diagnostic and therapeutic procedures* (n = 32,939/ 78.03%), followed by the *Operations on the cardiovascular system* (n = 24,623/ 58.33%) and the *Operations on the digestive system* (n = 9,885/ 23.42%). There are also 4,079 patients (9.66%) having *Procedures and interventions, not elsewhere classified.*

Table 4.3 Distribution of procedure groups. Note thata patient might receive more than one procedure and thus be included in more than one group. Percentage (%) is calculated over the total patients.

ICD code	Procedure label	n	%
00	Procedures and interventions, not elsewhere classified	4,079	9.66
01-05	Operations on the nervous system	5,105	12.09
06-07	Operations on the endocrine system	189	0.45
08-16	Operations on the eye	194	0.46
17	Other miscellaneous diagnostic and therapeutic procedures	37	0.09
18-20	Operations on the ear	56	0.13
21-29	Operations on the nose, mouth, and pharynx	700	1.66
30-34	Operations on the respiratory system	7,208	17.07
35-39	Operations on the cardiovascular system	24,623	58.33
40-41	Operations on the hemic and lymphatic system	1,239	2.94
42-54	Operations on the digestive system	9,885	23.42
55-59	Operations on the urinary system	1,066	2.53
60-64	Operations on the male genital organs	2,234	5.29
65-71	Operations on the female genital organs	286	0.68
72-75	Obstetrical procedures	52	0.12
76-84	Operations on the musculoskeletal system	3,465	8.21
85-86	Operations on the integumentary system	2,595	6.15
87-99	Miscellaneous diagnostic and therapeutic procedures	32,939	78.03

The distribution of diagnosis and procedure groups in Table 4.2 and Table 4.3 show that the variety of the data in the MIMIC-III database is wide. This means that this database, despite the limitation due to the anonymisation, is suitable for a range of healthcare studies.

4.1.6 Limitations of using MIMIC-III for process mining

The MIMIC-III database represents real healthcare data, with some limitations that it contains only the data of Critical Care patients and there is no direct access to the clinicians in the hospital. This leads to the limitation that the full analysis can only be done based on the available data. Despite this limitation, there are also some publications and websites describing this data that can be used to support analysis in this study.

Some identified limitations specific for process mining were omissions, incorrectness, incompleteness, and inaccuracy. Those can happen in different levels i.e. the events, case attributes, activity names or codes, timestamps, and attributes. This means that quality checking should be done thoroughly. This is because, for example, incorrectness could happen to an event, an activity name, a timestamp, or any other attribute. Another significant limitation is that all the dates have been shifted to future years (between 2100 and 2200) consistently for each patient to randomly distributed future dates. This means that analysis related to time between different patients, such as workflow analysis looking at busy days and the impact of bottlenecks e.g. of patients waiting for care on a busy day, cannot be deduced.

Despite these issues, the overall data quality of the MIMIC-III dataset was found to be good; there is a rich set of detailed event data covering a 10-year period and it comes from a representative of a real-life hospital. The MIMIC-III dataset still contains detailed information of real healthcare processes for individual patients during their time in the hospital including comprehensive details on administrative activities (admission, discharge, transfer to a ward, etc.) and clinical activities (triage, test and scans, diagnosis, etc.). Another reason why this dataset was suitable for this study was that there were a system change in the MIMIC-III dataset with a possibility to work out on separating patient admissions from those two different systems. It was a great opportunity for this study to compare processes in those two systems.

Section 4.1 has described the data characteristics, scope, data quality, representativeness, data variety, and the limitations of using MIMIC-III dataset in this study. Sections 4.2 and 4.3 go on to present two experiments using the MIMIC-III dataset. Section 4.2 presents the first experiment to apply process mining on the MIMIC-III dataset. The challenge of this experiment was on selecting the suitable tables to represent the patient pathways. Section 4.3 presents the second experiment to compare process in CV and MV as two subsequent systems. The challenge was to find a way to separate records in CV and MV and then compare the processes. Section 4.4 summarises this chapter.

4.2 Experiment 1: Process mining on the MIMIC-III dataset

This section presents the first experiment aimed to assess the suitability of the MIMIC-III dataset to be analysed using process mining approaches. A cohort of patients diagnosed with cancer was selected based on an initial discussion with UK-based clinical experts. Implementation of the main stages of the general methodology in this case study is presented in this section.

4.2.1 Stage 1: Planning and justification

Stage 1 was done by understanding the available data from the data descriptions on the official website and in papers related to MIMIC-III [162]. Historical data were generated from 16 event tables in MIMIC-III. There were no handmade models used in this study. The general research question was "*Can the MIMIC-III database be used for process mining in healthcare?*". This detailed research question was based on generic types of questions that are frequently posed by medical professionals in process mining projects, which are:

1) What are the most followed paths and what exceptional paths are followed?
2) Are there differences in care paths followed by different patient groups?
3) Where are the long waiting time activities in the process?

4.2.2 New stage: Database reconstruction

An additional stage in this study was database reconstruction. It was necessary to reconstruct the MIMIC-III dataset from the csv files to create a relational database in a local database management system (PostgreSQL). This stage included downloading the 26 csv files (6.2 GB in total) along with scripts to import the data into the PostgreSQL database.

Figure 4.3 presents the concept-level E-R diagram of the 26 tables in the MIMIC-III database. The approach of this study was to reconstruct the database to get the fullest possible dataset. The reconstructed database was then used for iterative extractions based on several criteria.

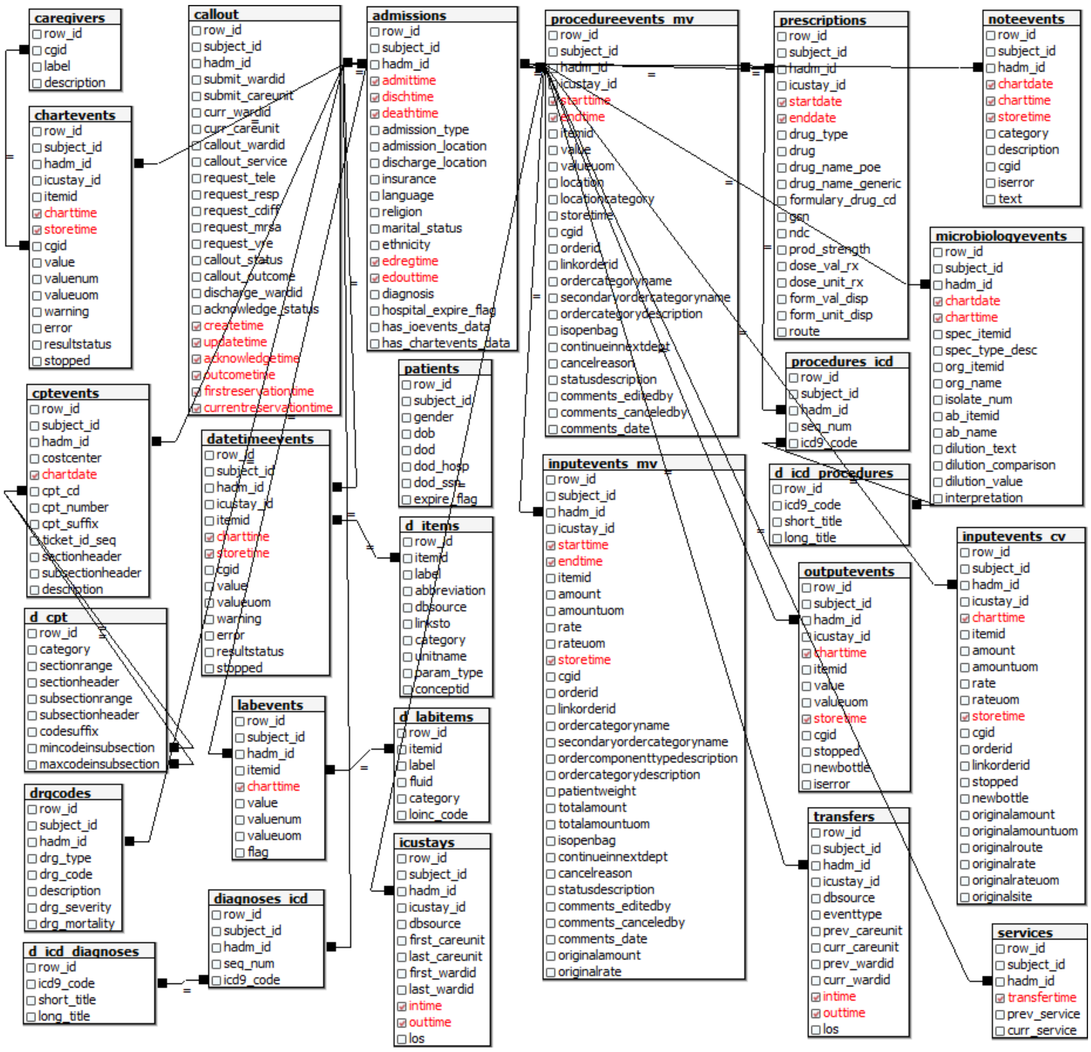


Figure 4.3 Entity-Relationship (E-R) diagram of the MIMIC-III database. The entities in red contain timestamped information which can be used to construct event log data for process mining.

4.2.3 Stage 2: Extraction, transformation, and loading

This stage started with selecting records of patients diagnosed with cancer. This selection was based on the ICD codes in the *diagnoses_icd* table for cancer diagnosis (140x-239x) [170]. Those patients were later grouped based on the cancer types, as presented in Table 4.2 in Section 4.1.5.

In total, 7,361 patients had at least one cancer diagnosis and were selected in this study. Those patients were then grouped based on the 13 types of cancer (see Table 4.4). The three largest groups are group 7, group 2, and group 8. The median age of patients in each cancer types ranges from 46 years (group 5) to 74 years (group 9). Median hospital length of stays (LOS) is 8 to 15 days, while the median ICU length of stay ranges from 2 to 3 days.

Table 4.4 Summary of cancer type in the MIMIC-III data

Type	Description (ICD9 codes in brackets)	a	b	c	d	e	f
1	Malignant neoplasm of lip, oral cavity, and pharynx (140-149)	87	135	135	65	8	3
2	Malignant neoplasm of digestive organs and peritoneum (150-159)	1,400	2,012	2,148	68	10	2
3	Malignant neoplasm of respiratory and intrathoracic organs (160-165)	1,056	1,540	1,561	69	8	2
4	Malignant neoplasm of bone, connective tissue, skin, and breast (170-175)	238	337	336	61	8	2
5	Kaposi's sarcoma (176)	14	19	20	46	8	2
6	Malignant neoplasm of genitourinary organs (179-189)	724	1,025	1,076	73	9	2
7	Malignant neoplasm of other and unspecified sites (190-199)	2,846	3,950	4,003	64	8	2
8	Malignant neoplasm of lymphatic and hematopoietic tissue (200-209)	1,110	1,692	1,876	62	14	3
9	Neuroendocrine tumours (209-209)	26	38	42	74	15	3
10	Benign neoplasm (210-229)	1,215	2,036	2,127	59	8	2
11	Carcinoma in situ (230-234)	45	65	66	70	13	2
12	Neoplasms of uncertain behaviour (235-238)	588	1,065	1,145	65	10	2
13	Neoplasms of uncertain nature (239)	60	105	108	67	9	2

Note: a = distinct patients, b = distinct admissions, c= distinct ICU stay id, d = median age (years), e = median hospital LOS (days), f = median ICU LOS (days)

Extraction of all cancer patient records was done by selecting records of cancer patients from each event table in MIMIC-III. For example, the query to extract cancer patient records in *chartevents* table is as follow.

```
SELECT
  c.subject_id, c.hadm_id, d.label, d.category, c.charttime
FROM
  Chartevents c, d_items d, diagnoses_icd di
WHERE
  c.itemid = d.itemid AND c.subject_id IN
  (SELECT DISTINCT subject_id FROM mimiciii.diagnoses_icd
  WHERE icd9_code BETWEEN '14%’ AND ‘24%');
```

All 16 event tables in the MIMIC-III database had been extracted using this method. They were then combined to create an *allevents* table. The three largest tables are *chartevents*, *labevents*, and *inputevents_cv*. A summary of the extracted records is presented in Table 4.5.

Table 4.5 Summary of table extracted

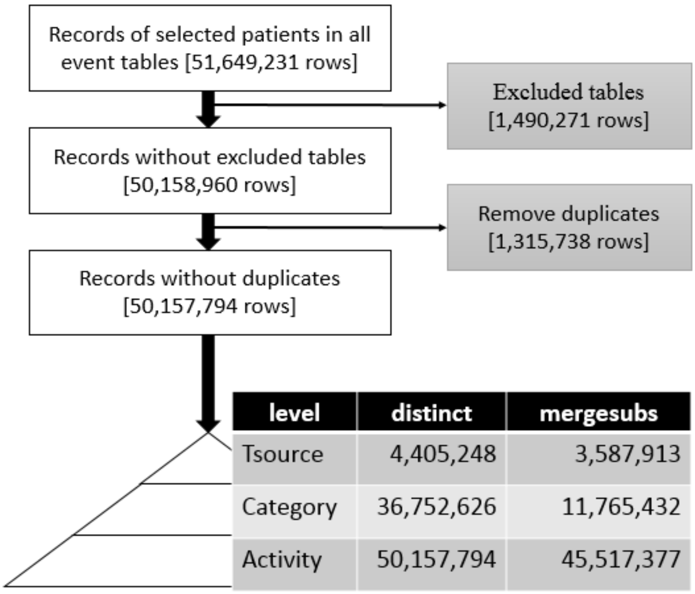
#	Table Name	Patients	Activities	Rows	Percentage
1	*admissions*	7,361	5	35,843	0.07
2	*callout*	4,771	6	27402	0.05
3	*chartevents*	7,359	2,580	**38,766,594**	**76.07**
4	*cptevents*	6,707	4	19,310	0.04
5	*datetimeevents*	5,648	148	925,542	1.82
6	*icustays*	7,345	2	22,976	0.05
7	*inputevents_cv*	3,924	756	**1,833,886**	**3.60**
8	*inputevents_mv*	3,850	251	664,209	1.30
9	*labevents*	7,351	556	**6,912,233**	**13.56**
10	*microbiologyevents*	3,553	47	11,670	0.02
11	*noteevents*	5,351	584	129,712	0.25
12	*outputevents*	7,278	415	824,665	1.62
13	*procedureevents_mv*	3,853	114	53,440	0.10
14	*prescriptions*	6,900	2,697	685,648	1.35
15	*services*	7,357	18	15,657	0.03
16	*transfers*	7,361	8	29,708	0.06
			allevents table	**51,649,231**	

The *allevents* table was created by combining all tables contained 51,649,231 rows, which would result in a ‘spaghetti’ model if directly used for process discovery. It is also important to mention that the original *allevents* table is dominated by *chartevents* (76.07%). Straightforward use of process mining will contain only *chartevents* and exclude other events. Further transformation of the dataset was then essential.

Transformation was performed in several steps, as shown in Figure 4.4. A transactional table was created consisting of *subject_id*, *activity*, *category*, *tname*, *charttime* and records were inserted from all tables extracted before.

Data processing was followed the steps in PM^2 methodology: (1) *Filtering log* – by excluding tables with a high percentage of missing data (*callout, cptevents*, and *prescription*) and handling duplicate records by keeping only one of them, (2) *Enriching log* –by creating three levels of details (table name, category, and activity label), (3) *Creating views* – this was done based on the level of detail needed in the next stage based on two general types of events recorded in the tables, administrative and clinical events, and (4) *Aggregating events* – this was done by applying “Merge subsequent events >> Merge taking first event” plugin in the ProM software.

A summary of those data transformation steps is shown in Figure 4.4.


Records of selected patients in all event tables [51,649,231 rows]
Excluded tables [1,490,271 rows]
Records without excluded tables [50,158,960 rows]
Remove duplicates [1,315,738 rows]
Records without duplicates [50,157,794 rows]

level	distinct	mergesubs
Tsource	4,405,248	3,587,913
Category	36,752,626	11,765,432
Activity	50,157,794	45,517,377

Figure 4.4 Data transformation. It includes log filtering, log enrichment, views creation, and event aggregation.

Loading, the final step in this stage was done by importing the file to ProM and processing it to discover process models using the specified algorithm/plugin. Unless otherwise stated, all plugins in ProM were applied with default parameter settings.

4.2.4 Stage 3: Mining and analysis

The next stage of this study was process mining and analytics. Process mining was done through process discovery, conformance checking, and enhancement. Process analytics was done to compare pathways of patient groups and analyse waiting times.

The general research question for this step was "*Can the MIMIC-III database be used for process mining in healthcare?*". The initial algorithm was the heuristics miner. Analysis and conformance checking were done to ensure that the discovered models represent realities in clinical settings. This was done by balancing the fitness, precision, and generalisation measures of the process model quality dimensions.

Figure 4.5 shows the trace variant diagram as one result. It shows the five most common variants out of 104 variants from 8,912 traces. The pathways represented the reality where patients could be admitted to standard hospital admission or registered through the ED, and could later be discharged from the hospital. The interesting pattern was that there were some patients discharged fully before actually being discharged from the ICU (*ICU out*) (see the fourth and fifth traces). This finding reflected the administrative process within the hospital.

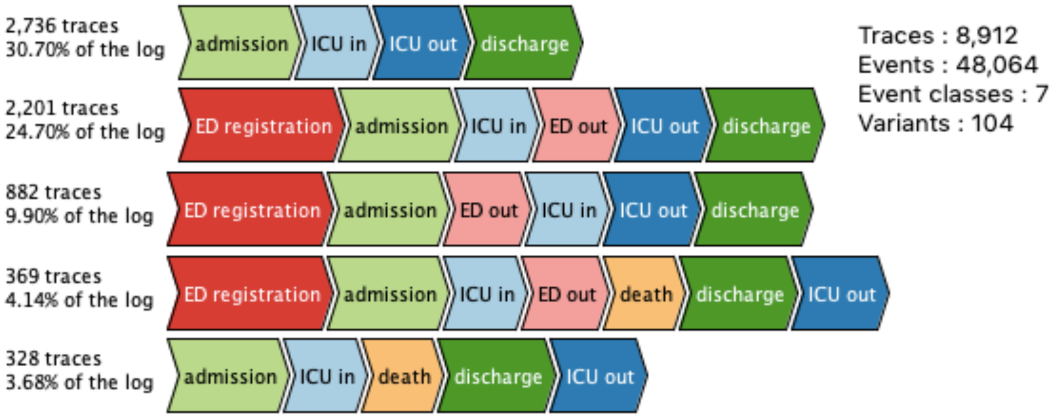


Figure 4.5 The five most common variants (from ProM). It shows the top five variants covering for >70% traces in the event log.

Figure 4.6 shows the discovered process model in the Business Process Modelling Notation (BPMN). This process model has a fitness score of 0.971, a precision score of 0.881 and a generalisation score of 0.989. All three conformance values were high, showing that the discovered model is indeed a good representation of the event log. A UK-based oncologist then reviewed the discovered models for sense-checking. One possible data quality issue was that *discharge* took place after *death* and *ICU out* took place after *discharge*. Those were found to reflect the hospital's standard administrative processes. There are also variations in the administrative steps as presented in Figure 4.6.

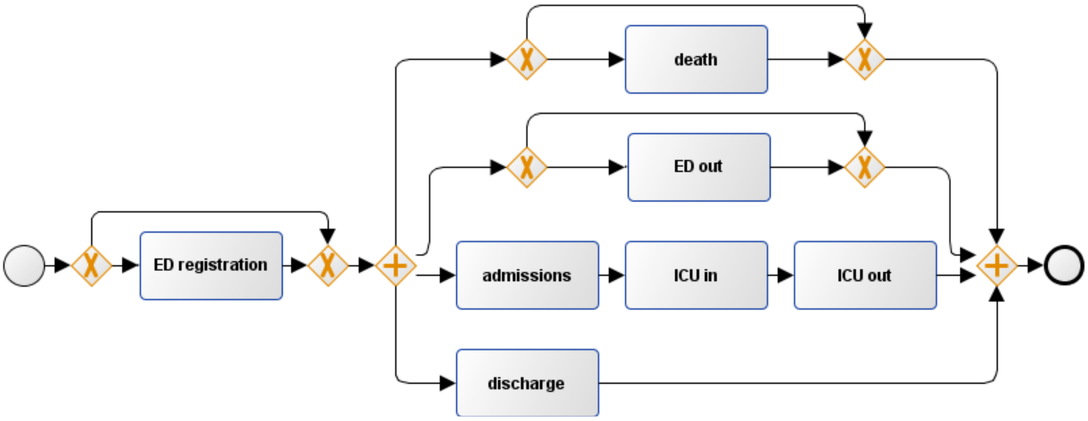


Figure 4.6 The BPMN process model of admissions and ICU stays (from ProM). This model shows variations in the administrative steps.

Following the same method, process models were created from each group to answer the research question "*Are there differences in care paths followed by different patient groups?*". The pathways of the three most common groups are shown in Figure 4.7.

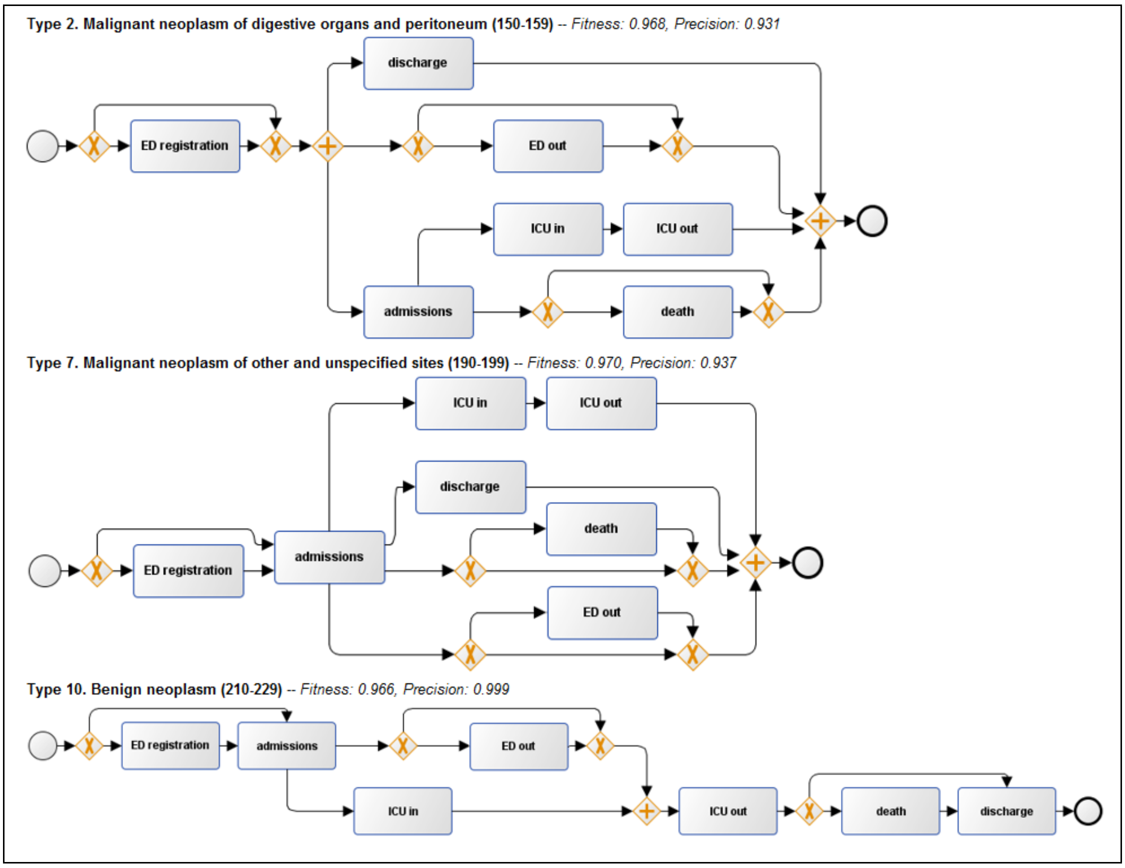

Type 2. Malignant neoplasm of digestive organs and peritoneum (150-159) -- Fitness: 0.968, Precision: 0.931
discharge
ED registration
ED out
ICU in
ICU out
admissions
death
Type 7. Malignant neoplasm of other and unspecified sites (190-199) -- Fitness: 0.970, Precision: 0.937
ICU in
ICU out
discharge
death
ED registration
admissions
ED out
Type 10. Benign neoplasm (210-229) -- Fitness: 0.966, Precision: 0.999
ED registration
admissions
ED out
ICU in
ICU out
death
discharge

Figure 4.7 Process models of three most frequent cancer types (from ProM). The process models show visual differences between patient groups.

The models of cancer type 2, type 7, and type 10 were reviewed by a UK clinical expert based on their visual utility. Some important findings are: (1) *ICU in, ICU out, admission* and *discharge* always happened in all three cancer types regardless of the sequence; (2) *ED registration, ED out* and *death* are possible events in all three types; (3) *admission* happened as the first event or after *ED registration* in all five types; and (4) *death* in type 10 is possibly happened only after ICU out, while it can happened before ICU out in the other four types.

Analysing differences of the pathways of different types of cancer can also be done by comparing conformance values, i.e. trace fitness, precision, and generalisation. The result is summarised in Table 4.6. Summary of the results shows that process models of all cancer types are representative of the traces, with average fitness of 0.843, average precision of 0.798, and average generalisation of 0.966. The minimum value is on the precision of cancer type 13 (0.692) and the maximum value is on the precision of cancer type 5 (1.000).

Table 4.6 Conformance values of each type of cancer

** F = fitness, P = precision, G = generalisation, MN = malignant neoplasm*

Type	Description	F	P	G
1	MN of lip, oral cavity, and pharynx (140-149)	0.813	0.767	0.961
2	MN of digestive organs and peritoneum (150-159)	0.813	0.816	0.977
3	MN of respiratory and intrathoracic organs (160-165)	0.830	0.826	0.981
4	MN of bone, connective tissue, skin, and breast (170-175)	0.819	0.822	0.952
5	Kaposi's sarcoma (176)	0.800	1.000	0.978
6	MN of genitourinary organs (179-189)	0.825	0.809	0.982
7	MN of other and unspecified sites (190-199)	0.835	0.822	0.991
8	MN of lymphatic and hematopoietic tissue (200-209)	0.894	0.716	0.995
9	Neuroendocrine tumours (209-209)	0.898	0.619	0.855
10	Benign neoplasm (210-229)	0.830	0.808	0.982
11	Carcinoma in situ (230-234)	0.882	0.859	0.975
12	Neoplasms of uncertain behavior (235-238)	0.849	0.813	0.978
13	Neoplasms of uncertain nature (239)	0.872	0.692	0.950
	average	**0.843**	**0.798**	**0.966**

Further analysis in this stage has been done to explore the waiting times for the admission pathways of all cancer patients, in order to answer research question "*Where are the long waiting time activities in the process?*". In this stage, the models have been extended by adding a time perspective.

The BPMN model was analysed with the "Replay a Log on Petri net for Performance/ Conformance Analysis" in ProM. The result was added on the original BPMN, as shown in Figure 4.8.

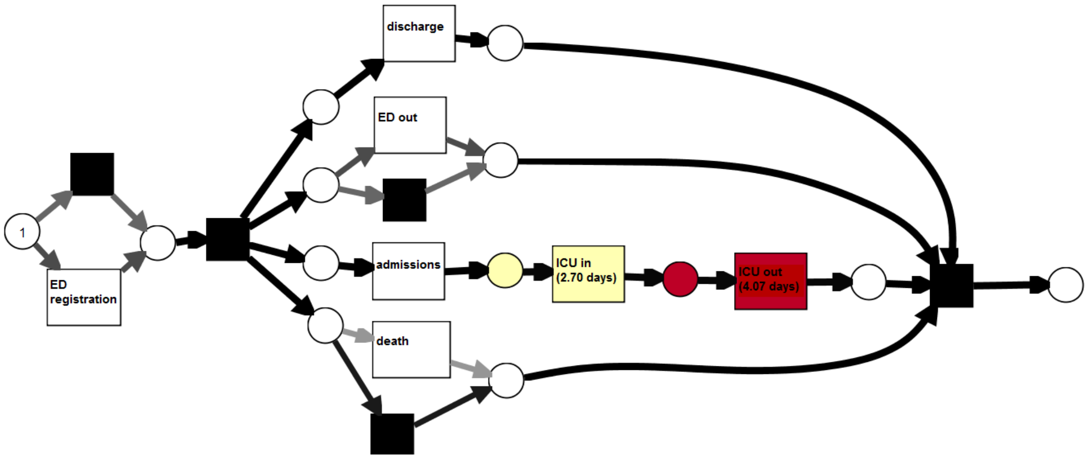


Figure 4.8 Waiting time analysis. This is a combined output from the *BPMN Analysis, Convert BPMN diagram to Petri Net,* and *Replay a Log on Petri net* plugins.

The analysis revealed that the longest waiting time is in *ICU out* (4.07 days on average), while the second longest waiting time is *ICU in* (2.70 days on average). The long waiting times in *ICU in* and *ICU out* give the insight to dig deeper into the lower level of activities between *ICU in* and *ICU out* to understand which activities contribute to the long waiting time in the ICU.

4.2.5 Stage 4: Evaluation

Evaluation of this experiment was done by analysing the suitability of the MIMIC-III dataset for process mining. The MIMIC-III database does not contain an event log, but it is possible to analyse the clinical processes in the MIMIC-III database with process mining approaches through the extraction of event data in 16 out of 26 tables within the MIMIC-III database. A new stage is needed to reconstruct the database in a database management system, such as MS SQL Server or PostgreSQL. The database reconstruction stage has been presented in Section 4.2.2. Evaluation of the findings was achieved by discussions with a UK-based oncologist. This approach has a quality concern that the UK-based oncologist might not have enough knowledge about the USA healthcare system. The task of the oncologist was only to do sense-checking and raise concerns on any possible issues related to the USA healthcare system.

An additional evaluation was done using 5-fold cross-validation. The event log was randomly partitioned into five equal-sized partitions (also called as folds). One fold was used as the validation data to test the model while the remaining four folds were used as the training data to build the process model. The cross-validation was done five times so that each of the five folds used exactly once as the validation data. The final results were estimated from the average of the five values of fitness, precision, and generalisation. The results of this evaluation are presented in Table 4.7. It shows that the process model represents the traces in the event log (high fitness) with relatively high precision and high generalisation.

Table 4.7 The results of five-fold cross-validation

#	Training folds	Validation fold	Fitness	Precision	Generalisation
1	2, 3, 4, 5	1	0.97024	0.6984	0.9873
2	1, 3, 4, 5	2	0.96156	0.8254	0.9794
3	1, 2, 4, 5	3	0.96961	0.9029	0.9838
4	1, 2, 3, 5	4	0.97389	0.6803	0.9873
5	1, 2, 3, 4	5	0.96792	0.8793	0.9784
average			**0.968644**	**0.79726**	**0.9832**

4.3 Experiment 2: Comparing CV and MV systems

This experiment aimed to identify the effects of the change in the EHR system from CareVue (CV) to MetaVision (MV). In this experiment, all patients in the MIMIC-III dataset were included. The first challenge in this experiment was to separate admissions from the CV and MV systems.

4.3.1 Stage 1: Planning and justification

Stage 1 (planning and justification) was done by understanding the dataset and identifying research questions. The research questions are:

1) Is it possible to create separated logs from CV and MV systems?
2) Can process mining be used to analyse differences in care paths in two systems?

Because all dates in the MIMIC-III have been shifted, a simple comparison of dates was not possible. However, there is a *db_source* column in the *d_items* table to identify the data source of each admission in the tables linked to the *d_items* table. This column will be used to create separated logs for CV and MV systems. In this experiment, process mining approach will be used in the model-based and log-based comparisons to analyse differences in care paths in the two systems.

4.3.2 Stage 2: Extraction, transformation, and loading

The checking on *itemid* column in *d_items* table is presented in Table 4.8. The differences between the two EHR systems can be identified through four tables: *chartevents, datetimeevents, inputevents*, and *outputevents*. The *dbsource* hospital would be ignored because our focus was on the CV and MV systems used.

Table 4.8 Details of itemid in d_item table

dbsource	linksto	occur
CareVue	*chartevents*	4,982
	datetimeevents	52
	inputevents_cv	2,929
	outputevents_cv	1,087
MetaVision	*chartevents*	924
	datetimeevents	141
	inputevents_mv	422
	outputevents_mv	74
	procedureevents_mv	125
hospital	*microbiologyevents*	436

This logic was used to insert a field in the *admissions* table that indicated which of the two systems had been used to record the admission. These new fields were used to ***extract*** the records in those tables and ***transform*** them into event logs of the CV and MV systems. The proportion of patients in the CV, MV, and both systems, is shown in Figure 4.9.

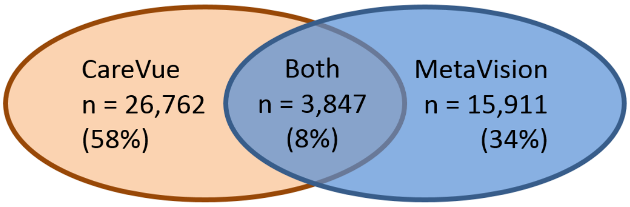


Figure 4.9 Proportion of CV, MV and both systems' patients. The 'both' group refers the number of patients who had records in both CV and MV systems.

As presented in Figure 4.9, the number of patients in the CV and MV only groups are not balanced. It is also important to note that within the data collection, the durations of the system usage were indeed imbalanced. The CV system was used for eight years (2001–2008) while the MV system was used for four years (2008–2012). A patient might have admissions in both the CV and MV systems. Only patients in the CV and MV groups were used.

The two separated event logs were then ***loaded*** to the process mining tools. Both logs compared to each other within the model-based and log-based comparison. The model-based comparison was done to compare process models discovered in those two systems. The log-based comparison was done to compare event logs in those two systems. Both model-based and log-based comparisons were done in *the mining and analysis (stage 3)* of this experiment.

4.3.3 Stage 3a: Model-based comparison

The model-based comparison started with discovering the process models of the separated event logs and followed by comparing those two process models. The process discovery step was done by loading the *allevents* table into the process mining tools to create a process map using several algorithms. Some tools used in this study were DISCO, ProM, and bupaR. The process model of all 16 events is not presented here because it is too complex.

As an illustration, Figure 4.10 shows the process model of five admission events of 26,762 patients in the CV system. There are 74,779 events in total for all patients, consisting of 30 variants. The most frequent variant is: *Admission →Discharge* (n= 9,273 / 40.2%). The median case duration is 7.4 days and the mean is 11 days. The second most frequent variant is *ED registration →Admission →ED out →Discharge* (n= 9,263 / 40.16%). The median case duration is 7.4 days and the mean is 10.6 days. These two variants are the two most frequent ones with and without the ED episode.

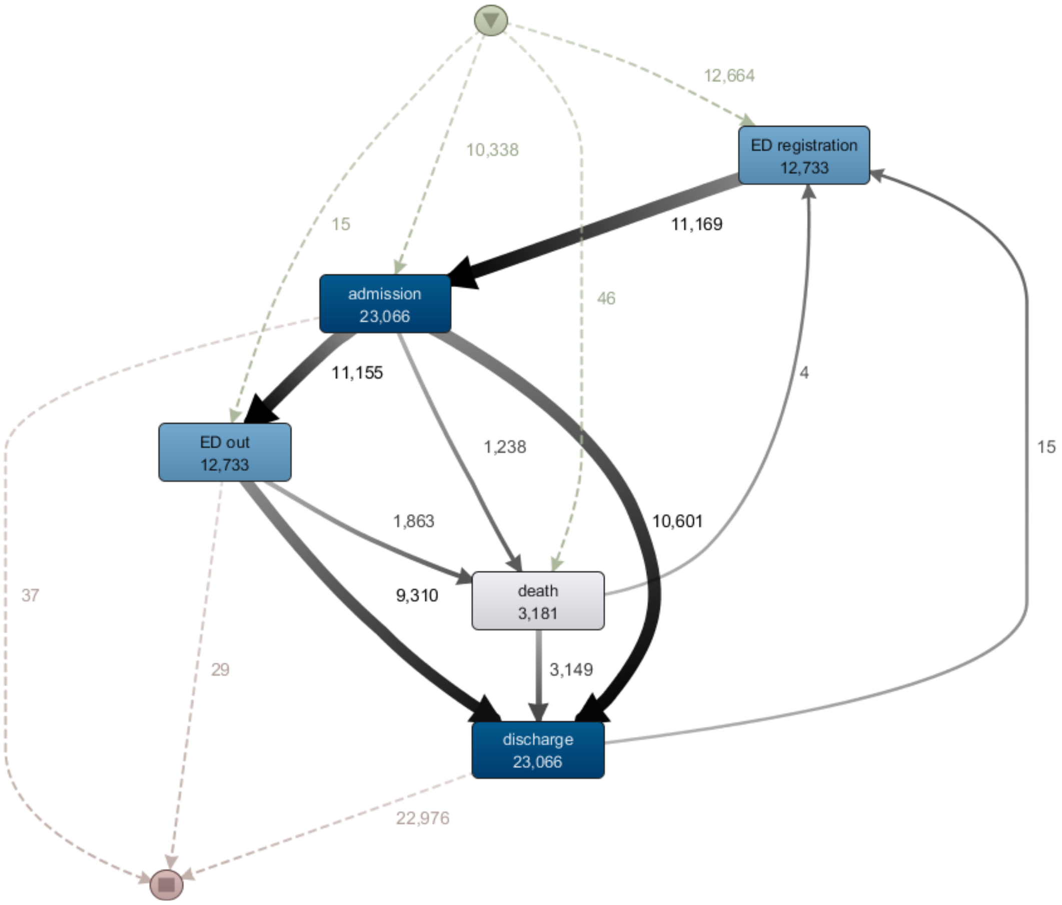


Figure 4.10 Process model of admissions in the CV system showing 70% of the most common paths (from DISCO)

The process models discovered from the CV and MV event logs were then compared using the DifferenceGraph plugin in ProM 5.2. This plugin identifies differences and commonalities between two process models. The result is presented in Figure 4.11.

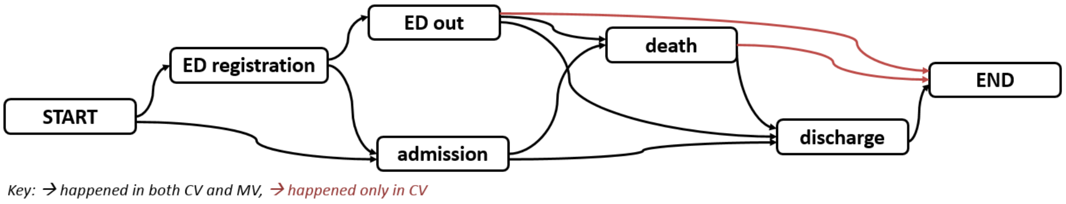


Figure 4.11 DifferenceGraph of admissions in CV and MV resulted from ProM. The red arcs represent the differences in the two event logs.

Figure 4.11 shows the differences between admissions in the CV and MV systems, which found changes in the last activities that happened within those two systems. The admissions in CV ended with either *discharge* (35,788 or 99.874%), *death* (26 or 0.073%), or *ED out* (19 or 0.053%). But all *admissions* in MV ended with *discharge* (19,623 or 100%), suggesting that MV has better consistency on the administrative records than CV.

The CV process model gave a fitness score of 0.9996, a precision score of 0.8784, and a generalisation score of 0.9006; while the MV process model gave a fitness score of 0.9990, a precision score of 0.9382, and a generalisation score of 0.8916. This suggested that both models can replay the observed behaviour (high fitness), describe the system generally (high generalisation), and not allow for too much divergent behaviour (high precision). The combination of interpreting event frequencies, the DifferenceGraph, and the conformance measures leads to the conclusion that the EHR system change did affect the process model and quality and required further investigation. More results of model-based comparison is presented in Appendix C.3. The limitation of this approach is that it is depended on the visual difference of the models. Future improvements are needed to improve this approach.

4.3.4 Stage 3b: Log-based comparison

Log-based comparison of patient data in CV and MV systems was done using log profiling and process comparator approaches. The grouping into CV and MV admissions were done based on the method described in Section 4.3.1. Log profiling was done for each table in the MIMIC-III database.

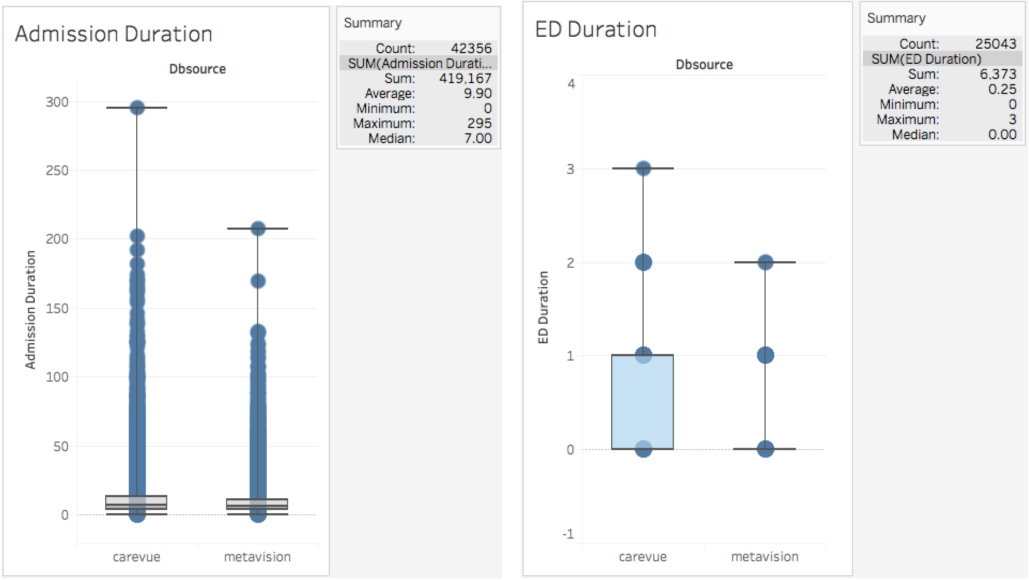
For example, from the admissions table, some important information is:

- NEWBORN admissions only happened in CV, and are excluded in the experiment. The proportion is presented in Table 4.9.

Table 4.9 Proportion comparison of CV and MV

Database	Admission type	Patients	Admissions	% Admissions
CareVue	*EMERGENCY*	15,499	18,508	80.24
	ELECTIVE	3,433	3,594	15.58
	URGENT	941	964	4.18
MetaVision	*EMERGENCY*	13,492	16,112	83.53
	ELECTIVE	2,831	2,955	15.32
	URGENT	223	223	1.16

- The average admission duration in MV (8.97 days) is shorter than CV (10.67 days).
- There were data quality issues detected in CV where ED duration < 0. After non-valid data were excluded, the average ED duration in MV (0.23 days) was shorter than CV (0.28 days), as presented in Figure 4.12.



Summary

Count:	42356
SUM(Admission Durati...	
Sum:	419,167
Average:	9.90
Minimum:	0
Maximum:	295
Median:	7.00

Summary

Count:	25043
SUM(ED Duration)	
Sum:	6,373
Average:	0.25
Minimum:	0
Maximum:	3
Median:	0.00

Figure 4.12 Comparison of hospital admission of CV and MV patients on admission durations (left) and ED durations (right).

Figure 4.12 shows that both the average of admission durations and the average of ED durations in MV are shorter than in CV. The results of log profiling from the other tables are documented in detail in Appendix C.2.

Log-based comparison [149] aims to detect relevant differences between processes based on what was recorded in event logs. In this study, the log-based comparison was done using the Process Comparator plugin in ProM [150]. This plugin compares two event logs: CV as the first log and MV as the second log. The traces in those two event logs are merged to create the reference log. Process metrics in this experiment is trace frequency, with an alpha significant level of 5%. The significant difference between the two logs is presented as a state transition diagram annotated with colours based on the oracle of the effect size. The colour legend is shown in Figure 4.13.

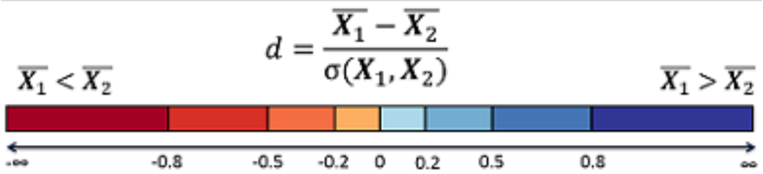

$d = \frac{\overline{X_1} - \overline{X_2}}{\sigma(X_1, X_2)}$
$\overline{X_1} < \overline{X_2}$
$\overline{X_1} > \overline{X_2}$
-∞
-0.8
-0.5
-0.2
0
0.2
0.5
0.8
∞

Figure 4.13 Colour legend in the Process Comparator plugin (from ProM). Darker colours represent higher significance differences. Red represents lower average in group 1 than in group 2; blue represents higher average in group 1 than in group 2.

The log-based comparison was done per table to compare processes in the CV and MV systems. For example, the log-based comparison was done to the admissions in the CV and MV systems. Characteristics of the admissions in the two systems are presented in Table 4.10.

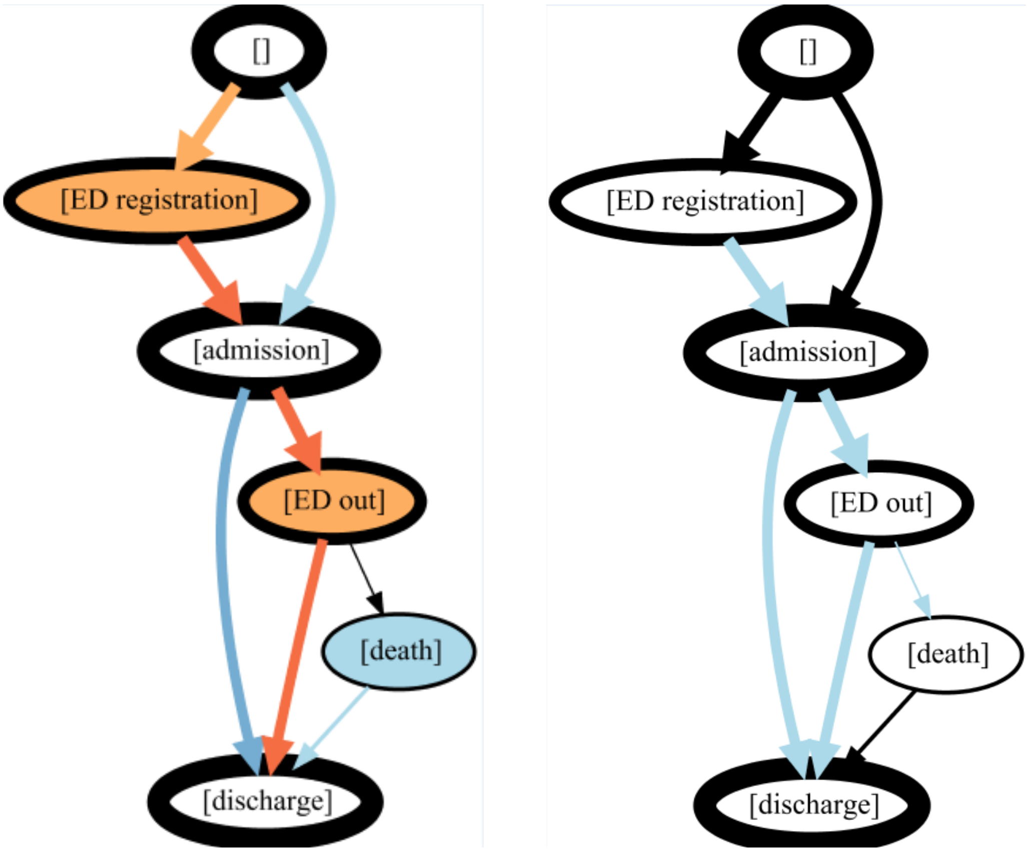
Table 4.10 Characteristics of the admissions in CV and MV systems

Characteristics	CV	MV
Cases [hadm_id]	23,066	19,290
Events	74,779	65,213
Event classes	5	5
Events per case *	3 [2-5]	3 [2-5]
Event classes per case *	3 [2-5]	3 [2-5]
Variants	30	20
Pair-wise difference	71.43%	

** average [min – max]*

Table 4.10 shows that admissions in the CV system are larger than the MV system. The number of event classes (5), events per case (3 [2-5]) and event classes per case (3 [2-5]) are the same in both systems. The number of variants in CV (30) is larger than in MV (20), suggesting that the variability in the older system is higher than in the new system. The Process Comparator plugin shows that the pairwise difference is 71.43%.

The detailed differences are shown in Figure 4.14. The comparison diagram in Figure 4.14(a) shows that the frequency of patients having ED registration and ED out in MV are larger than CV, while the frequency of patients with death records in CV is larger than MV. Figure 4.14(b) shows that the trace durations are similar in both systems, with slightly longer duration in CV than MV.



(a) metric: trace frequency

(b) metric: duration

Figure 4.14 Comparison diagrams of hospital admission of CV and MV patients. The trace frequency metric shows more significant differences than the duration metric.

A technical challenge in doing this comparison is that the processing time of the Process Comparator plugin in ProM 6.8 is high. Some tables could not be compared using this plugin because of the processing time. There are only six tables that can be compared using the process comparator plugin. Those are: *admissions*, *callout*, *labevents, noteevents*, *outputevents*, and *services*. Of those six tables, the *noteevents* table has the highest difference (92.31%) between the CV and MV systems. This was followed by *outputevents* (87.50%) and *services* (78.75%). The highest difference in the *noteevents* is apparently due to the highly different number of event classes in CV (21) and MV (1165). The number of events per case and variants is consequently high. The *outputevents* both in CV and MV have 12 event classes, but the number of events per case and variants in those two systems are highly different. The *services* have 19 event classes in CV and 17 event classes in MV. The high difference is apparently because the average number of events per case in both systems is one, meaning that most patients only have one event in *services*. The results are summarised in Table 4.11.

Table 4.11 Log-based comparison summary

Tablename	DB	Cases	Events	Event classes	Events per case*	Var	Diff.(%)
admissions	CV	23,066	74,779	5	3 [2-5]	30	71.43
	MV	19,290	65,213	5	3 [2-5]	20	
callout	CV	7,037	32,635	4	5 [3-23]	85	50
	MV	15,698	89,245	6	6 [3-35]	366	
labevents	CV	22,763	1,744,263	6	77 [1-300,203]	21,108	72,73
	MV	19,176	878,686	6	46 [1-76,333]	16,774	
noteevents	CV	22,081	347,882	21	16 [1-931]	244	92.31
	MV	7,331	175,715	1165	24 [1-701]	6,724	
outputevents	CV	21,915	1,781,202	12	81 [1-5208]	943	87.50
	MV	18,697	1,228,067	12	66 [1-2104]	3,352	
services	CV	23,045	29,530	19	1 [1-7]	484	78.75
	MV	19,278	24,810	17	1 [1-9]	505	

** average [min – max]*

4.3.5 Stage 4: Evaluation

Evaluation of this experiment has been done by analysing the suitability of the methods used to compare processes in two different systems. The source of the data collected in the MIMIC-III was changed in 2008. A backward approach was done to mark hospital admissions with the data source from which they had been recorded, which are the CV and MV systems.

Comparison of the processes in those two data sources (CV in 2001–2008 and MV in 2008–2012) was conducted to explore the effects of a system change to the treatment process. The comparison was done based on the process model and log conformance. The model-based comparison on the administrative events in CV and MV found some interesting results, such as that *discharge* is recorded in all patients in MV, but not in CV. This suggested that the system change has improved the consistency of the data recording. The log-based comparison with the Process Comparator plugin in ProM resulted in the percentage of difference between two logs at a time. This method is robust and based on a statistical test that can be adjusted as required. The examples of the resulted comparison diagrams in Figure 4.14 are easy to understand and showing the differences of two logs based on trace frequency and duration of the process. The limitation of using the process comparator plugin was that this method was technically limited and only applicable to five out of 11 tables in this experiment.

This suggested a potential research experiment to create a new method for log-based comparison. This is explored further with the PPM Cancer data in Chapter 6.

4.4 Summary

This chapter has described the analysis of the MIMIC-III dataset. The assessment of the data quality has been published in a journal paper in 2018 [168] and cited in 7 articles. Some parts of experiment 1 have been published in a conference paper in 2018 [169] and cited in 7 articles.

This case study can be analysed based on different perspectives, as described in Section 1.2. From the health service perspective, process mining has been able to answer frequency posed questions, such as those in the first experiment in Section 4.2. Health service professionals can get some insights into the most followed paths and the exceptional paths, the differences in care paths followed by different patient groups, and the long waiting time activities in the process. From the process mining perspective, this case study has been shown how MIMIC-III can be used for process mining in healthcare. Some limitations have been discussed, but the experiments have shown that process discovery and conformance checking can be applied to analyse MIMIC-III dataset. From the information system perspective, experiment 2 in Section 4.3 concluded an important lesson learned that a system change affected the data recorded in the system. Limited discussion of the clinical perspective in this case study is due to the limited access to the data source of the MIMIC-III database.

Chapter 5

Case study 2: Experiments using the PPM Chemotherapy Data

Analysis of the MIMIC-III dataset as the first case study was presented in Chapter 4. This chapter presents the experiments using the Patient Pathway Manager (PPM) Chemotherapy dataset as the second case study. Section 5.1 presents the data description of the PPM Chemotherapy dataset. Section 5.2 and Section 5.3 present the two experiments in this case study. Section 5.2 has already been presented in a poster in the NCRI Cancer Conference 2018 in Glasgow, UK, entitled “Process mining to explore variation in chemotherapy pathways for breast cancer patients”. The abstract of the poster has been published in the British Journal of Cancer supplement [171].

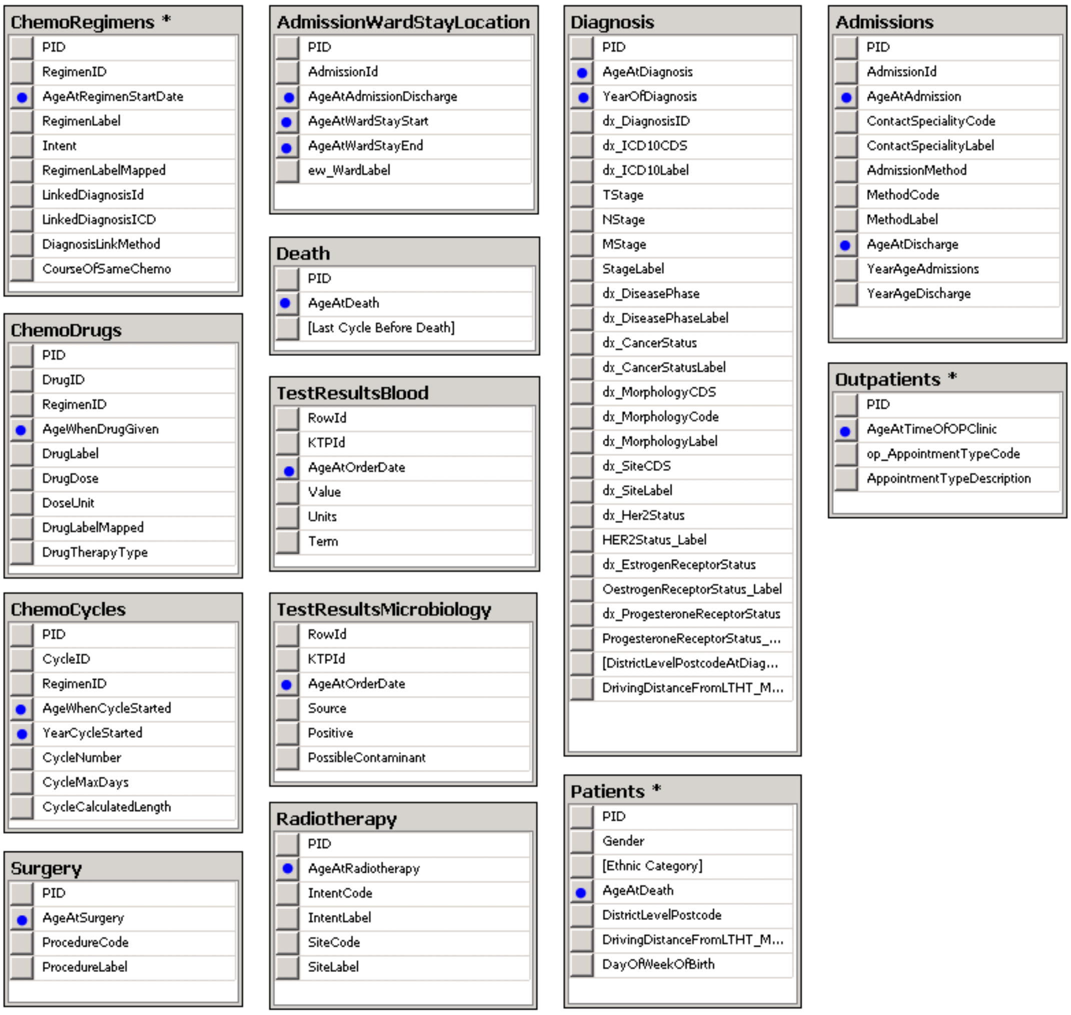
5.1 Data description

The PPM Chemotherapy dataset was introduced in Section 1.4.2 and described in Section 3.4.2. This section explores this dataset in more detail.

5.1.1 Data characterisation

The PPM Chemotherapy dataset is a pseudonymised dataset transferred from the LTHT to a secure SQL database on a virtual machine at the University of Leeds (UoL). The data has been checked, cleaned, and aggregated before approval for transfer to the research team. The previous study used the PPM Chemotherapy data to define a predictive model identifying patients receiving cancer treatment who were at the highest risk of developing life-threatening complications, such as neutropenia.

There are 13 event tables in the PPM Chemotherapy dataset, as presented in Figure 5.1. All tables are related based on the Patient Identification (PID), except the *TestResultsBlood* and the *TestResultsMicrobiology* tables that are linked based on the Knowledge Transfer Partnership ID (KTPId) and were extracted from the Result database. The KTPId was the identifier used in the previous study using this dataset.



ChemoRegimens *
- PID
- RegimenID
- ● AgeAtRegimenStartDate
- RegimenLabel
- Intent
- RegimenLabelMapped
- LinkedDiagnosisId
- LinkedDiagnosisICD
- DiagnosisLinkMethod
- CourseOfSameChemo

ChemoDrugs
- PID
- DrugID
- RegimenID
- ● AgeWhenDrugGiven
- DrugLabel
- DrugDose
- DoseUnit
- DrugLabelMapped
- DrugTherapyType

ChemoCycles
- PID
- CycleID
- RegimenID
- ● AgeWhenCycleStarted
- ● YearCycleStarted
- CycleNumber
- CycleMaxDays
- CycleCalculatedLength

Surgery
- PID
- ● AgeAtSurgery
- ProcedureCode
- ProcedureLabel

AdmissionWardStayLocation
- PID
- AdmissionId
- ● AgeAtAdmissionDischarge
- ● AgeAtWardStayStart
- ● AgeAtWardStayEnd
- ew_WardLabel

Death
- PID
- ● AgeAtDeath
- [Last Cycle Before Death]

TestResultsBlood
- RowId
- KTPId
- ● AgeAtOrderDate
- Value
- Units
- Term

TestResultsMicrobiology
- RowId
- KTPId
- ● AgeAtOrderDate
- Source
- Positive
- PossibleContaminant

Radiotherapy
- PID
- ● AgeAtRadiotherapy
- IntentCode
- IntentLabel
- SiteCode
- SiteLabel

Diagnosis
- PID
- ● AgeAtDiagnosis
- ● YearOfDiagnosis
- dx_DiagnosisID
- dx_ICD10CDS
- dx_ICD10Label
- TStage
- NStage
- MStage
- StageLabel
- dx_DiseasePhase
- dx_DiseasePhaseLabel
- dx_CancerStatus
- dx_CancerStatusLabel
- dx_MorphologyCDS
- dx_MorphologyCode
- dx_MorphologyLabel
- dx_SiteCDS
- dx_SiteLabel
- dx_Her2Status
- HER2Status_Label
- dx_EstrogenReceptorStatus
- OestrogenReceptorStatus_Label
- dx_ProgesteroneReceptorStatus
- ProgesteroneReceptorStatus_...
- [DistrictLevelPostcodeAtDiag...
- DrivingDistanceFromLTHT_M...

Patients *
- PID
- Gender
- [Ethnic Category]
- ● AgeAtDeath
- DistrictLevelPostcode
- DrivingDistanceFromLTHT_M...
- DayOfWeekOfBirth

Admissions
- PID
- AdmissionId
- ● AgeAtAdmission
- ContactSpecialityCode
- ContactSpecialityLabel
- AdmissionMethod
- MethodCode
- MethodLabel
- ● AgeAtDischarge
- YearAgeAdmissions
- YearAgeDischarge

Outpatients *
- PID
- ● AgeAtTimeOfOPClinic
- op_AppointmentTypeCode
- AppointmentTypeDescription

Figure 5.1 The database structure of the PPM Chemotherapy data. Blue dot marks the *Age* attribute that later be used to create a generated timestamp.

5.1.2 Scope

The scope of the dataset for this study can be described as follows. This dataset contains the routinely collected data of cancer patients (ICD-10 codes C00–C97, D00–D48) who were first diagnosed between 2004 and 2012, and treated with chemotherapy, in the Leeds Cancer Centre. As of 13 August 2013 (PPM Chemotherapy data collection), there were 1.76M patients in the PPM EHR system with 92,044 of them having received chemotherapy or radiotherapy. The PPM EHR system has been used since 2003 to hold the clinical records of all cancer patients including their diagnosis, treatment and outcome treated at the LTHT.

The data included patient demographics (age, gender, ethnicity, distance from LTHT), details of chemotherapy drugs and doses, details of the cancer diagnosis and related co-morbidity, results of blood tests, results of microbiological investigations, details of acute admissions and outpatient reviews. Further discussions with clinical experts suggested that the selection of events included in the PPM Chemotherapy dataset are suitable in analysing the chemotherapy pathways in cancer patients. Specific scope was suggested to focus on breast cancer patients receiving EC-90 as an adjuvant chemotherapy.

5.1.3 Data quality

Data quality of the PPM Chemotherapy dataset was assessed based on the Weiskopf & Weng framework [133]. The completeness was checked through element presence checking, the correctness and plausibility were checked through validity checking, and the concordance through element agreement. Each of the data quality dimensions is presented in the following paragraphs.

The ***completeness*** of the data was checked based on the element presence checking of the three minimum required attributes for process mining, which are *case_id*, *activity*, and *timestamp*. There were no timestamps available in the dataset, but *Age* (in the number of days) was recorded for each activity. Those *Age* columns were then used to generate dates by assigning '01-01-2020' as the date of birth for all patients. A summary of the element presence checking is presented in Table 5.1.

Table 5.1 Element presence checking of PPM Chemotherapy dataset

Table source	Element presence		
	case_id [%]	Activity [%]	Time (AgeAt-) [%]
Admissions	PID [91]	AdmissionMethod [91]	Admission, Discharge [91]
AdmissionWard_ StayLocation	PID [91]	*WardLabel+Start [91] *WardLabel+End [91]	WardStayStart [91] WardStayEnd [91]
ChemoCycles	PID [92]	Cycle Number [92]	CycleStarted [92]
ChemoDrugs	PID [99]	Drug Label [99]	DrugGiven [99]
ChemoRegimens	PID [99]	RegimenLabel [99]	RegimenStartDate [99]
Death	PID [56]	'Death' [56]	Death [56]
Diagnosis	PID [98]	dx_ICD10Label [98]	Diagnosis [64]
Outpatients	PID [92]	AppointmentTypeDesc [91]	TimeOfOPClinic [92]
Patients	PID [100]	**'Death' [36]	Death [36]
Radiotherapy	PID [53]	IntentLabel+ SiteCode [53]	Radiotherapy [53]
Surgery	PID [68]	ProcedureLabel	Surgery [68]
TestResultsBlood	PID [93]	Term [93]	OrderDate [93]
TestResultsMicrobiology	PID [50]	Source [50%]	OrderDate [50]

** Activity names are created based on available recorded Age (in number of days)*

*** 'Patients' table contains only Death events, duplicated but Death' table is complete*

The ***correctness*** of the PPM Chemotherapy dataset was examined through a validity check. Some doubts had been discussed with the domain experts and those have been validated, such as: (1) Chemotherapy cycle numbers were recorded with a range of 1 to 93; (2) Cycle max days were recorded with a range of −7 to 547; and (3) Year of diagnosis ranged from 1921 to 2015. The recorded year of diagnosis 1921 suggested a data quality problem and required a filtering step to include only a reasonable range of diagnosis year. Some others were excluded from the data, such as: (1) Age at discharge = NULL; (2) Four cases ended with Admissions; (3) Age at Radiotherapy = −13959; and (4) Six missing values in ProcedureLabel. All those findings represent incorrect records in the data. This condition happened because, when the PPM database had begun to be used to record data, older records were manually typed in from the paper-based records.

The ***concordance*** of the PPM Chemotherapy dataset was checked through element agreement. Some important findings are: (1) the *ChemoCycles*, *ChemoDrugs*, and *ChemoRegimens* tables contain different numbers of patients, as should be the case. A discussion with clinical experts revealed that the *ChemoCycles* table is the most reliable among those three tables. (2) 'Death' events were recorded in two tables: the *Death* table and the *Patients* table. The records in the *Death* table are more than those in the *Patients* table.

The ***plausibility*** was checked through validity checking. The PPM Chemotherapy dataset contains the records of 31,511 patients diagnosed with cancer between 1921 and 2012, of whom 29,009 had received chemotherapy, 21,395 surgery and 16,792 radiotherapy in Leeds. From a discussion with a domain expert in the early stages of this study, we had confirmed that these numbers did not represent the whole number of cancer patients. This is because the data collection was focused on the patients receiving chemotherapy only and was not meant to reflect the reality where there should be more patients receiving radiotherapy than those receiving chemotherapy.

5.1.4 Representativeness

The representativeness of the PPM Chemotherapy dataset is related to the PPM Cancer data, the database where the PPM Chemotherapy dataset was extracted from. The PPM Cancer data was extracted from the PPM EHR system in the LTHT. The Leeds Cancer Centre in the LTHT is one of the largest specialist cancer centres in the UK [21]. Based on this fact, the data of cancer patient treatment in the PPM Cancer dataset can be argued to be representative of the data of cancer patient treatment in the UK. Additional analysis of the representativeness was done by checking if the most common cancers in the UK can also be found as the common cancers in the PPM Cancer dataset.

In the UK, there were 303,135 new cancer diagnoses registered in 2016. Over half (52.7%) of all registrations were either one of four cancer types (ICD-10 codes): breast cancer (C50), prostate cancer (C61), lung cancer (C34), and colorectal cancer (C18–C20) [172]. The PPM Chemotherapy dataset consists of the data of patients who had at least one diagnosis of cancer between 1921 and 2015. Table 5.2 shows the occurrence of those four most common cancer types in the PPM Chemotherapy dataset.

Table 5.2 Representativeness of PPM Chemotherapy dataset

Cancer diagnosis	Occurrence	Patients (%)	Rank
Breast cancer (C50)	5,616	3,952 (32.6%)	**2**
Prostate cancer (C61)	1,116	901 (7.4%)	**10**
Lung cancer (C34)	4,286	3,205 (26.4%)	**3**
Colorectal cancer (C18-C20)	6,151	4,418 (36.4%)	**1**
All four cancers	**17,169**	**12,131**	

Table 5.2 summarises number of patients in the PPM Chemotherapy dataset for the four most common types of cancer. The differences in the rank of the most common types of cancer is possibly because the PPM Chemotherapy dataset contains the data of patients treated with chemotherapy only. It is also possible that the patients are different, especially due to the referral hospital aspect. Based on an article on the Cancer Research UK website [173], the treatments for patients diagnosed with prostate cancer are radiotherapy (30%), surgery (15%), and chemotherapy (3%). Based on this analysis, the PPM Chemotherapy dataset is representative of colorectal cancer, breast cancer and lung cancer, but not for prostate cancer patients.

5.1.5 Data variety

The data variety of the PPM Chemotherapy dataset can be explained by exploring the diagnosis groups. The diagnosis in the PPM Chemotherapy dataset was recorded based on the ICD-10 diagnosis codes. The distribution of the diagnosis groups is presented in Table 5.3.

Table 5.3 Diagnosis group distribution in PPM Chemotherapy data

#	Description	Patients
1	(C00–C14) Malignant neoplasms, lip, oral cavity and pharynx	1,418
2	(C15–C26) Malignant neoplasms, digestive organs	**8,914**
3	(C30–C39) Malignant neoplasms, respiratory system and intrathoracic organs	3,563
4	(C40–C41) Malignant neoplasms, bone and articular cartilage	268
5	(C43–C44) Malignant neoplasms, skin	1,326
6	(C45–C49) Malignant neoplasms, connective and soft tissue	966
7	(C50–C58) Malignant neoplasms, breast and female genital organs	**6,318**
8	(C60–C63) Malignant neoplasms of male genital organs	1,745
9	(C64–C68) Malignant neoplasms, urinary organs	1,832
10	(C69–C72) Malignant neoplasms, eye, brain and central nervous system	835
11	(C73–C75) Malignant neoplasms, endocrine glands and related structures	153
12	(C76–C80) Malignant neoplasms, secondary and ill-defined	749
13	(C81–C96) Malignant neoplasms, stated or presumed to be primary, of lymphoid, haematopoietic and related tissue	**4,248**
14	(C97) Malignant neoplasms of independent (primary) multiple sites	0
15	(D00–D09) In situ neoplasms	803
16	(D10–D36) Benign neoplasms	1,778
17	(D37–D48) Neoplasms of uncertain or unknown behaviour	370

The PPM Chemotherapy dataset contains all cancer diagnosis groups, except on Malignant neoplasms of independent (primary) multiple sites (C97). Of all 31,511 patients recorded in the PPM Chemotherapy dataset, the three most common diagnoses are the cancer of digestive organs (8,914 patients/ 28%), breast and female genital organs (6,318 patients/ 20%), and lymphoid, haematopoietic and related tissue (4,248 patients/ 13%).

5.1.6 Limitations of using the PPM Chemotherapy dataset

The PPM Chemotherapy dataset was used for analysing cancer treatment pathways in the Leeds Cancer Centre. Some limitations identified in this study are in its representativeness, anonymisation on dates, and data granularity.

The *representativeness* limitation was due to the specific focus on data extraction. The PPM Chemotherapy dataset consists only of the clinical data of cancer patients receiving chemotherapy. Chemotherapy is one of three common treatments for cancer patients (other than surgery and radiotherapy). This means that the dataset consists only a part of the total number of cancer patients and is not representative of the whole population of cancer patients in the LTHT.

The *anonymisation* limitation was due to the approach taken to de-identified dates in the database. The dates were replaced with a calculated age of the patients (in the number of days since date of birth) when they experienced the events. In this study, this is a limitation because one of the required attributes needed in process mining is the timestamp. This limitation has been approached by creating generated dates. This makes it possible to analyse activity sequences for each patient, but not for cross-patient analysis, such as workload and bottleneck analysis.

The *granularity* limitation was that all the generated dates provide only analysis measured down to the day and there was no way the order of activities that happened on the same day could be further defined. This was not an issue in analysing chemotherapy cycles because it is not possible for a patient to get two cycles within the same day. This issue became a problem when analysing chemotherapy combined with other events, such as pathology and investigation.

5.2 Experiment 3: Process mining the PPM Chemotherapy dataset

The PPM Chemotherapy dataset was used previously to develop risk algorithms in oncology. This dataset contains non-identifiable data of chemotherapy patients treated in the LTHT. This experiment aims to apply process mining approach to analyse the variability of the treatment process in the PPM Chemotherapy dataset. The analysis was focused on chemotherapy treatment. The initial challenge was that the dataset has been de-identified, where all dates were transformed into the patient age in the number of days. The sequence of activities is reliable, but analysis across patients such as examining busy time was not possible.

5.2.1 Stage 1: Planning and justification

This experiment aimed to reproduce the previous research [34] using the same data. The contribution of this experiment was to improve the previous study by providing a structured method for pathways analysis using process mining approaches. The primary research question was "*Can process mining be used to analyse the variability of treatment in the PPM Chemotherapy dataset?*".

The table structures of the PPM Chemotherapy dataset have been presented in Figure 5.1. The real year was available in two tables: *ChemoCycles* and *Diagnosis*. Chemo cycle years were recorded from 1996 to 2015. The experiment of the PPM Chemotherapy dataset description included the steps to identify elements of the event log from the PPM Chemotherapy dataset. The experiment documentation of the PPM Chemotherapy dataset description is presented in Appendix D.1.

5.2.2 Stage 2: Extraction, transformation, and loading

The *extraction* of the dataset for this experiment was done to select all events related to all eligible patients. Patients were included if they had (i) a diagnosis of metastatic breast cancer (ICD-10 C50) and received adjuvant epirubicin and cyclophosphamide (EC-90) chemotherapy or (ii) colorectal cancer (ICD-10 C18–C20) and received palliative oxaliplatin and infusional 5-fluorouracil chemotherapy. Those selection criteria followed the previous study. The selected cohort consists of 738 breast cancer patients and 418 colorectal cancer patients, a total of 1156 patients.

The three basic elements of the dataset required for process mining are *case id, activity name,* and *timestamp*. A summary of those three basic elements in the PPM Chemotherapy dataset is as follows:

- **Case ID**s are available in all tables: *PID/KTPId*, *AdmissionId*, or *RegimenId*. This study used PID/KTPId to analyse the pathways of patients. PID is the patient identification from the PPM database and KTPId is the patient identification from the Result database.
- **Activity name** can be derived from each table except for *Patients* table. For example, the activity name from the *Admissions* table is *admission*, while the activity names from the *ChemoCycles* table are *Cycle-[number of cycles]*.
- There is no **timestamp** identified, but there is the possibility to obtain the sequence based on *Age* (in the number of days) and Years in the tables.

The event log can be created by combining all records in those identified in the event tables. The details of the *allevents* table are presented in Table 5.4. The *allevents* table contains more than 21 million records, with the three most frequent activities being *TestResultsBlood (83.54%)*, *Outpatients (4.56%)*, and *Admissions (3.73%)*. As shown in Table 5.4, the database is dominated by *TestResultsBlood* due to the focus of the data collection for the previous research related to blood test results of patients receiving chemotherapy. In this study, the focus is on analysing the pathways of patients receiving chemotherapy and not on the test results.

Table 5.4 Details of the allevent table

#	Source	Activity Class	Patients	Total Rows	%
1	*Admissions*	2	28,609	794,863	3.73
2	*AdmissionWardStayLocation*	360	28,609	469,886	2.20
3	*ChemoCycles*	93	29,009	198,096	0.93
4	*ChemoDrugs*	301	31,421	706,670	3.31
5	*ChemoRegimens*	2945	31,354	67,707	0.32
6	*Death*	1	17,701	17,701	0.08
7	*Diagnosis*	653	30,753	56,123	0.26
8	*Outpatients*	69	28,878	973,177	4.56
9	*Radiotherapy*	541	16,792	31,703	0.15
10	*Surgery*	1345	21,395	74,889	0.35
11	*TestResultsBlood*	34	29,151	17,814,931	83.54
12	*TestResultsMicrobiology*	23	15,678	118,947	0.56
Total				**21,324,693**	100

A direct use of the *allevents* table to create process models would result in a spaghetti model and would not be understandable. Some alternatives to analyse the process are to select events of interest or to analyse specific cohorts of patients. This study focused on the first approach to select events of interest.

Data *transformation* was done following the Process Mining Project Methodology (PM^2) approach, which includes *creating views, aggregating events, enriching logs,* and *filtering logs.* These were done with ProM 6.7 plugins. A summary of the data processing is as follows:

- Change the order of events having the same timestamp. This was important because the timestamp data are in days. The order of events was set to follow the logical order of the events based on a discussion with clinical experts.
- Merge subsequent events. This was done to merge some routine test results, including neutropenia and bacteraemia. Only the first occurrence of the event was considered when an event repeated in sequence.

- Rename/merge events as required. For example, *Elective admission (S2)* → *Chemotherapy (S1) = Chemotherapy (S1)*. This reflected the condition where the chemotherapy event was also recorded as an elective admission.
- Add artificial START and END events. Each trace was added by a START before the first event and an END after the last event. This was needed because the current ProM plugins cannot deal with multiple start and end events, so that it is important to ensure that all traces have exactly one START and one END event.

Additional work was done by following the activity coding done in the previous study [34]. The list of activity codes, names, and their descriptions is presented in Table 5.5.

Table 5.5 Activity codes and descriptions of the *allevent* table

Code	Activity	Description
S1	*Chemotherapy*	Chemotherapy was delivered
D0	*Home discharge*	Home discharge after chemotherapy
S2	*Elective admission*	Hospital attendance other than chemotherapy
F1	*Admission*	Other type of admissions
D7	*Neutropenic*	GP contact with neutropenia (neutrophil count <1.5x10^9/L)
S6	*Bacteraemia*	Bacteraemia (positive blood culture) during admission
D3	*Urgent outpatient*	Urgent review in outpatient
D6	*Death*	Death during hospital stay

5.2.3 Stage 3: Mining and analysis

This stage included process mining and process analytics. The process discovery was done to the selected events, using the Interactive Data-aware Heuristic Miner (iDHM) plugin in ProM. The iDHM plugin allows several representations to be chosen, including the directly-follows graph, dependency graph, causal net, or Petri Net.

1) Process mining of the general events

Process discovery resulted in a directly-follows graph as presented in Figure 5.2. A directly-follows graph describes what activities follow one another directly, based on the records in the event log. It was chosen because this graph provides important information visually, including START and END events, colours representing activity frequencies and arcs with the number of patients following the paths.

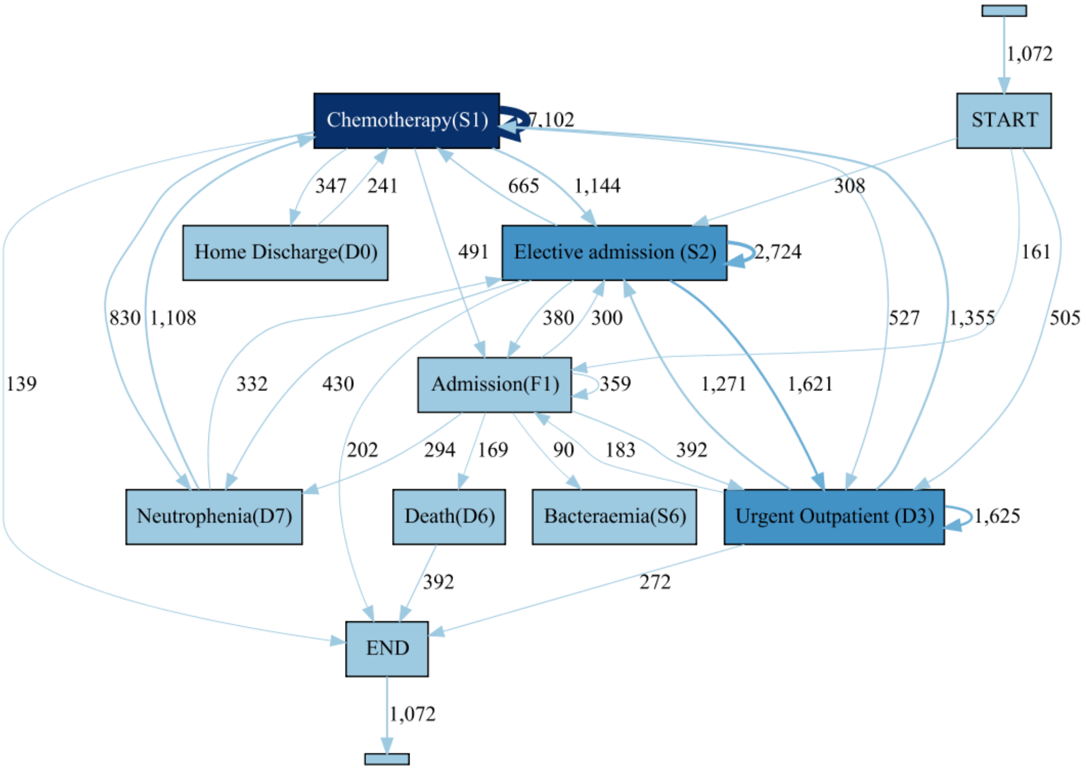


Figure 5.2 Directly-follows graph the PPM Chemotherapy dataset. This is a result from the interactive Data-aware Heuristics Miner (iDHM) plugin in ProM.

Figure 5.2 shows that patient pathways are started with either an *Elective admission, Admission,* or an *Urgent outpatient*. Patients from an *Elective admission* or *Urgent outpatient* might go for a round of *Chemotherapy* and *Home discharge* repeatedly until this is followed by either a *Neutropenia, Admission, Elective admission,* or *Urgent outpatient*. Following an *Admission*, patients might also have a *Neutropenia, Bacteraemia, Elective admission, Urgent outpatient,* or *Death*.

2) *Process mining of the chemotherapy cycles*

Further analysis was done to focus on the events during the chemotherapy cycles of breast cancer patients receiving EC-90 chemotherapy (n= 738). This was done by analysing a lower-level event, which is the chemotherapy cycle number. The results were a process model, a trace variant, and a dotted chart shown in Figure 5.3-5.5.

It is evidenced in the process model in Figure 5.3 that the number of patients from one cycle to the next is decreasing. The median duration between chemotherapy cycles is 21 days, which reflects the common duration in the treatment.

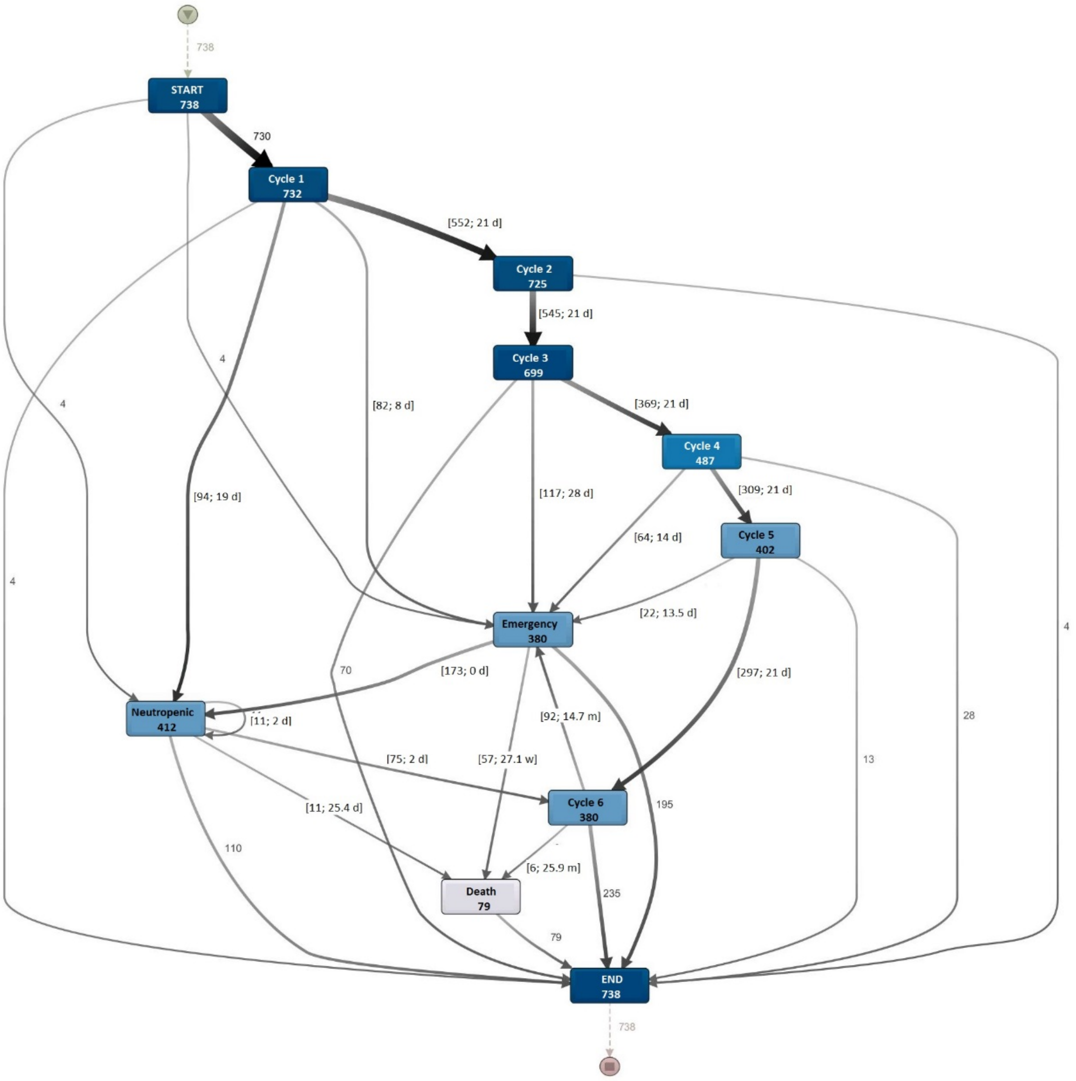


Figure 5.3 Process model showing 70% of the most common pathways. It was built in DISCO. The number below an activity name shows the number of patients undertaking that activity. The number on arc shows the detail of the flow [number of patients; median duration].

Figure 5.4 shows the six most common trace variants followed by more than ten patients. The most common trace variant is a sequence of Cycle 1 to Cycle 6 (n=120; 16.3%), followed by a sequence of Cycle 1 to Cycle 3 (n=56; 7.6%), and a sequence of Cycle 1 to Cycle 3 followed by an Emergency admission (n=37; 5%). It is shown that Emergency was mostly happened after Cycle 3 (see Variant 3) or Cycle 6 (see Variant 4), while Neutropenia happened after Cycle 5 (see Variant 6).

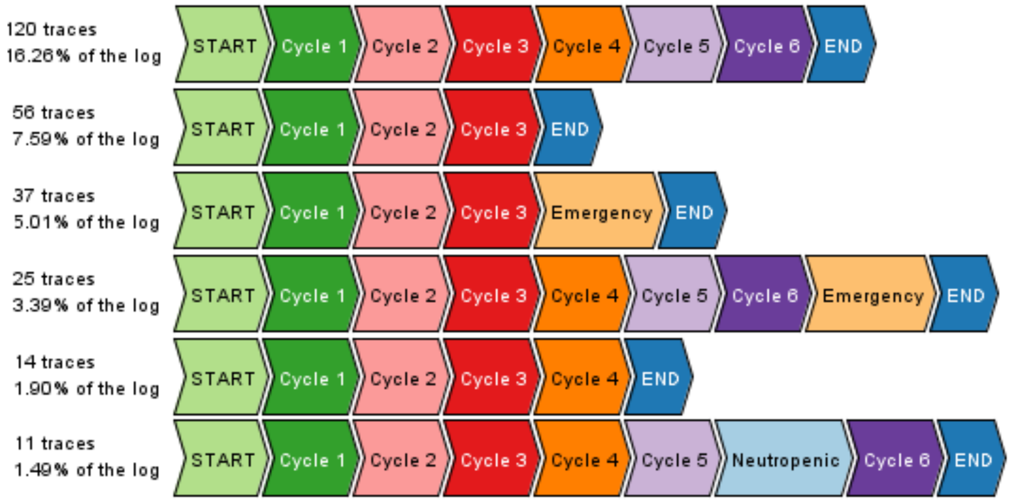


Figure 5.4 The six most common trace variants of chemotherapy cycles. Activities included are Cycle 1 through Cycle 6, Neutropenia, and Emergency.

Figure 5.5 shows a dotted chart of routine chemotherapy cycles of patients treatments of up to 7 years. The diagram shows groups of patients who had not completed six cycles of chemotherapy, who completed six cycles of chemotherapy, and who had more complicated courses of treatment. In total 51% (n=376) of patients did complete all six cycles, with 21% of the total (n=158) without any acute event while 30% (n=218) had at least one acute event including emergency admission or neutropenic sepsis. Of the 49% (n=362) patients who did not complete six cycles, 28% (n=207) had acute events with the remainder 21% (n=155) not completing for other reasons.

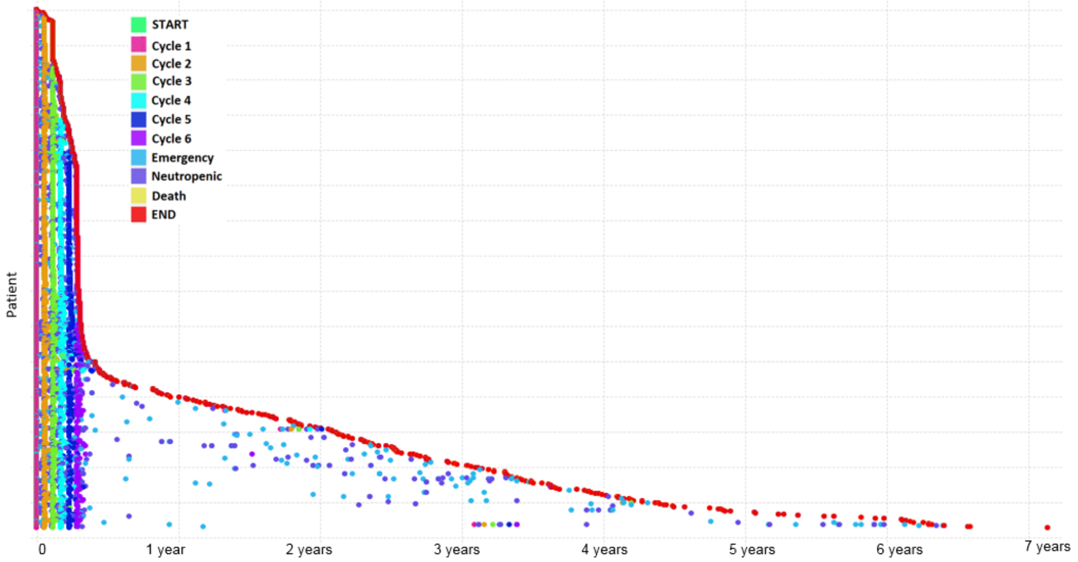


Figure 5.5 Dotted chart showing patient pathways over treatment duration. The x-axis shows duration from the first activity**.** The y-axis shows patient id, sorted from the shortest to the longest durations.

3) Process analytics: conformance-based process change over time

Another analysis was done to identify process change over time. Conformance checking was done by analysing the conformance values of the process model. A reference process model has been presented in Figure 5.2 with a fitness of 0.826, a precision of 0.342, and a generalisation of 0.999, suggesting that the model did represent the traces in the event log, was generalisable, but not very precise.

The analysis was enhanced to examine the process changes over time. It was done by checking the conformance of the event log per year to the reference process model.

Table 5.6 Fitness, precision, and generalisation over years

Year	*Events*	*Cases*	*Variants (%)*	*Fitness*	*Precision*	*Generalisation*
2004	2,555	20	20 (100)	0.7828	0.3112	0.9500
2005	4,363	32	31 (97)	0.8843	0.2784	0.9832
2006	3,155	29	29 (100)	0.8745	0.3296	0.9755
2007	4,083	44	43 (98)	0.8743	0.3029	0.9927
2008	13,082	143	136 (95)	0.7484	0.3385	0.9986
2009	14,889	173	154 (89)	0.7354	0.3459	0.9981
2010	13,781	147	138 (94)	0.7447	0.3327	0.9986
2011	14,255	168	138 (82)	0.7451	0.3669	0.9989
2012	13,377	138	118 (86)	0.7568	0.3583	0.9963
2013	14,302	149	125 (84)	0.7711	0.3719	0.9962
2014	21,803	24	24 (100)	0.7622	0.3894	0.9632
			Average	**0.78905**	**0.33869**	**0.98647**

The results are displayed in Table 5.6. It shows that the number of cases each year ranges from a minimum of 20 in 2004 to a maximum of 173 in 2009. The trend of the increasing number of cases from 2004 until 2013 was discussed with clinical experts and was said to conform the reality. This was related to the fact that the hospital is a large cancer centre with a growing capacity of cancer treatment. The fitness ranges from 0.7354 to 0.8843 (average = 0.7891). The precision ranges from 0.2784 to 0.3894 (average = 0.3387). The generalisation ranges from 0.95 to 0.9989 (average = 0.9865). The trend of those conformance values is presented in Figure 5.6.

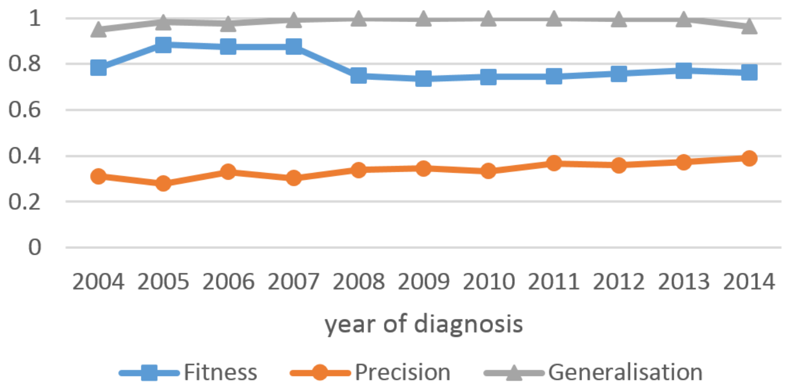


Figure 5.6 Trend of trace fitness, precision, and generalisation over time. The generalisation was generally steady, while precision and fitness fluctuated over time.

Figure 5.6 shows that the precision and fitness are not steady over time, suggesting there must be some changes. There is also a trade-off between fitness and precision where a more precise model would have lower fitness. There are two periods where the process had potentially changed, which are 2004-2005 and 2007-2008.

5.2.4 Stage 4: Evaluation

The evaluation was done through discussions with clinical experts. Those included a researcher of the previous study using the same dataset, a database administrator of the de-identified dataset, and a senior oncologist working with the PPM database. Discussions were done to get a description of the previous study, identify potential gaps in the research, and evaluate both intermediate and final results.

Discussion on the early stages was done to plan and justify the experiment. It was also done to discuss the decisions on the extraction, transformation, and loading steps. Results of this early discussion include the order of the same timestamp and merging events. The clinical expert provided the logical order of events happened on the same day. Merging events were also decided based on a suggestion by the clinical expert to filter out duplicated records in the dataset.

General comments of the intermediate and final results are that process mining is potentially useful to improve the previous study generating process model of chemotherapy treatment. The analysis is particularly helpful to visualise the variations from standard chemotherapy pathways, including incomplete treatment and adverse events. The conformance-based comparison was also useful for analysing process change over time, but the metrics need to be improved to identify change points reliably. The detected change points were likely related to the introduction of the PPM EHR system in 2003 and the change from PPM1 to the current PPM in 2007.

One limitation was that the preprocessing relied on the selection of the best set of events. It requires a good understanding of the dataset and a specific cohort of patients. This experiment was started by analysing the higher-level process as presented in Figure 5.2. It was then followed by examining the chemotherapy cycles including the adverse events, as shown in Figure 5.3- Figure 5.5. The decision to analyse the more specific pathway can be seen as a way to select the best set of events of interest. The more detailed analysis of the chemotherapy pathways is presented in Appendix D.2.

5.3 Experiment 4: Trace clustering for similarity analysis

The most obvious challenge of doing process mining in healthcare data is handling the high number of trace variants. One way to solve this is to apply trace clustering to cluster similar traces into one grouping. The experiment was done with hierarchical clustering using edit distance in the PPM Chemotherapy dataset. The complete step by step of this experiment is presented in this section.

5.3.1 Stage 1: Planning and justification

This stage was done to plan and justify the experiment. This experiment aims to improve the analysis of the variability of treatment in the PPM Chemotherapy dataset. The primary research question was "*Is it possible to use trace clustering to analyse the variability of treatment in the PPM Chemotherapy dataset?*". This is an additional experiment to support the previous experiment with a specific focus on exploring the potential to use hierarchical trace clustering method. The hypothesis was that the hierarchical clustering could be used to group similar traces and create separated process models for each cluster. This experiment used Python and DISCO.

5.3.2 Stage 2: Extraction, transformation, and loading

General stages in this experiment followed those done in Experiment 3 as presented in Section 5.2. The additional steps were:

(1) coding activity names,

(2) creating an edit distance matrix,

(3) hierarchical trace clustering,

(4) determining the optimal number of clusters,

(5) retrieving clusters, and

(6) describing process models of the clusters.

Step (1) was part of Stage 2 and is described in the following paragraph. Step (2) to step (6) were parts of Stage 3 and are described in Section 5.3.3.

Coding activity names was done to simplify the activity names into a character for the edit distance matrix creation. The coding step was done using pandas library in Python. Traces were hierarchically clustered using the *scipy.cluster.hierarchy* library. The optimal number of clusters was determined using the silhouette method. The activity names are presented in Table 5.7. The four most frequent activities are *Chemotherapy (n=1,072 / 100%), Day case review (n=798/ 74.4%), Home discharge (700/ 65.3%),* and *Urgent outpatient (n=664/ 61.9%).*

Table 5.7 Coding activity name, the occurrence, and the number of patients

Code	Activity Name	Occurrence (%)	n (%)
A	Chemotherapy (S1)	5,910 (23.5)	1,072 (100)
B	Urgent outpatient (D3)	1,944 (7.7)	664 (61.9)
C	Non-neutropenic (S5)	1,192 (4.7)	603 (56.3)
D	Day case review (D4)	7,179 (28.6)	798 (74.4)
E	Death (D6)	388 (1.5)	388 (36.2)
F	Admission non-neutropenic (S4)	1,308 (5.2)	541 (50.5)
G	Home discharge (D0)	1,928 (7.7)	700 (65.3)
H	Neutropenic (D7)	1,107 (4.4)	511 (47.7)
I	Not bacteraemia (S8)	884 (3.5)	425 (39.6)
J	Elective admission (S2)	618 (2.5)	360 (33.6)
K	Admission progressing neutropenic (S9)	41 (0.2)	38 (3.5)
L	Bacteraemia (S6)	143 (0.6)	108 (10.1)
M	Admission neutropenic (S7)	324 (1.3)	218 (20.3)

5.3.3 Stage 3: Mining and analysis

The process mining step of the complete event log resulted in a process model, as presented in Figure 5.7. The fitness score was 0.826 and the precision score was 0.342. The median duration was 24.4 months and the mean duration was 28.9 months. The number of events per patient ranged from 1 to 220, with a mean of 21. There were 1,042 variants out of 1,072 patients, suggesting that this process had high variability and there was a need to find groups of patients having similar traces to create simpler and more understandable process models.

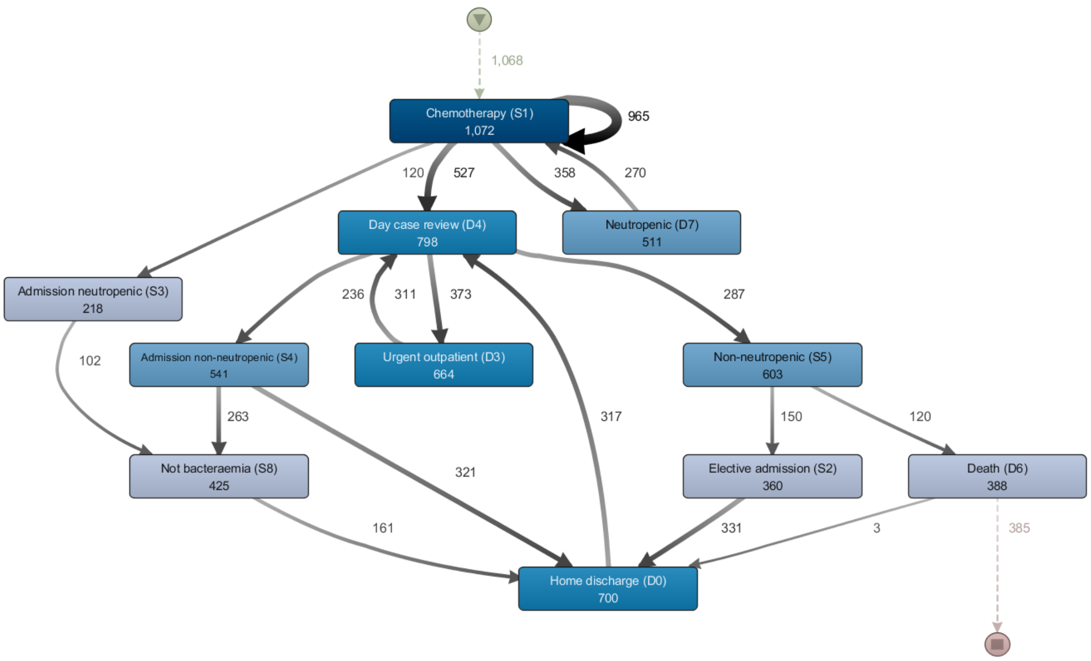


Figure 5.7 General process model resulted from DISCO with 80% activities and 0% paths

In the analysis step, the edit distance matrix was created using the Levenshtein distance [174]. It was originally used to measure the distance between two words as the minimum number of single-character edits (insertions, deletions, or substitutions) required to change one word into the other. This method is suitable for this experiment because the Levenshtein distance is a string metric for measuring the difference between two sequences. A trace with the coded activity names created a sequence of letters, thus comparing those sequences is the same as comparing words. The code to create a distance matrix is as follow.

```
#create distance matrix
import editdistance
import numpy as np
np.set_printoptions(threshold = np.inf)
def f(x,y):
    return editdistance.eval(x,y)

def cartesian_product(*arrays):
    la = len(arrays)
    dtype = np.result_type(*arrays)
    arr = np.empty([len(a) for a in arrays] + [la], dtype=dtype)
    for i, a in enumerate(np.ix_(*arrays)):
        arr[...,i] = a
    return arr.reshape(-1, la)

v=np.vectorize(f)
arr = cartesian_product(cva.activity, cva.activity).T
arr = v(arr[0, :], arr[1, :])
```

```
dist_df = pd.DataFrame(arr.reshape(-1, cva.shape[0]),
index=cva.index, columns=cva.index)
dist_df
```

The output is a distance matrix, as follow.

```
trace  1    2    3    4    5    6    7    8    9    10  ...   1070  1071  1072
trace                                                   ...
1         0    2    2    4    6    5    2    4   16   35 ...     88     8   26
2         2    0    1    2    8    3    4    3   14   33 ...     87     6   24
3         2    1    0    3    8    4    4    2   15   34 ...     86     6   25
4         4    2    3    0   10    5    6    2   13   32 ...     86     5   22
5         6    8    8   10    0    8    4   10   22   38 ...     94    14   32
6         5    3    4    5    8    0    6    6   17   30 ...     87     8   25
7         2    4    4    6    4    6    0    6   18   36 ...     90    10   28
8         4    3    2    2   10    6    6    0   13   33 ...     84     6   24
9        16   14   15   13   22   17   18   13    0   27 ...     73    12   17
10       35   33   34   32   38   30   36   33   27    0 ...     63    29   21
…
[1072 rows x 1072 columns]
```

The next step was to use the edit distance matrix to create a hierarchical clustering using *scipy.cluster.hierarchy*. This library clustered data based on agglomerative clustering [175], where objects are clustered based on their similarity. Pairs of objects with high similarity were merged iteratively until all objects are combined in one big cluster. The result can be presented as a tree-based representation of the objects, named a *dendrogram*.

Traces were hierarchically clustered using the scipy.cluster.hierarchy library. The code of the hierarchical clustering is as follow.

```
#hierarchical clustering
import scipy.cluster.hierarchy as hcl
from scipy.spatial.distance import squareform
from scipy.cluster.hierarchy import dendrogram
import matplotlib.pyplot as plt

linkage = hcl.linkage(dist_df, method='ward')
#ward minimises the sum squared distances within all clusters
plt.figure(figsize=(20, 10))
plt.title('Hierarchical Clustering Dendrogram')
plt.xlabel('sample index')
plt.ylabel('distance')
dendrogram(
    linkage,
    truncate_mode='lastp',  # show only the last p merged clusters
    p=10,  # show only the last p merged clusters
    leaf_rotation=90.,  # rotates the x axis labels
    leaf_font_size=8.,  # font size for the x axis labels
    show_contracted=True
```

The resulting dendrogram from this experiment is shown in Figure 5.8, which groups 1,072 traces hierarchically.

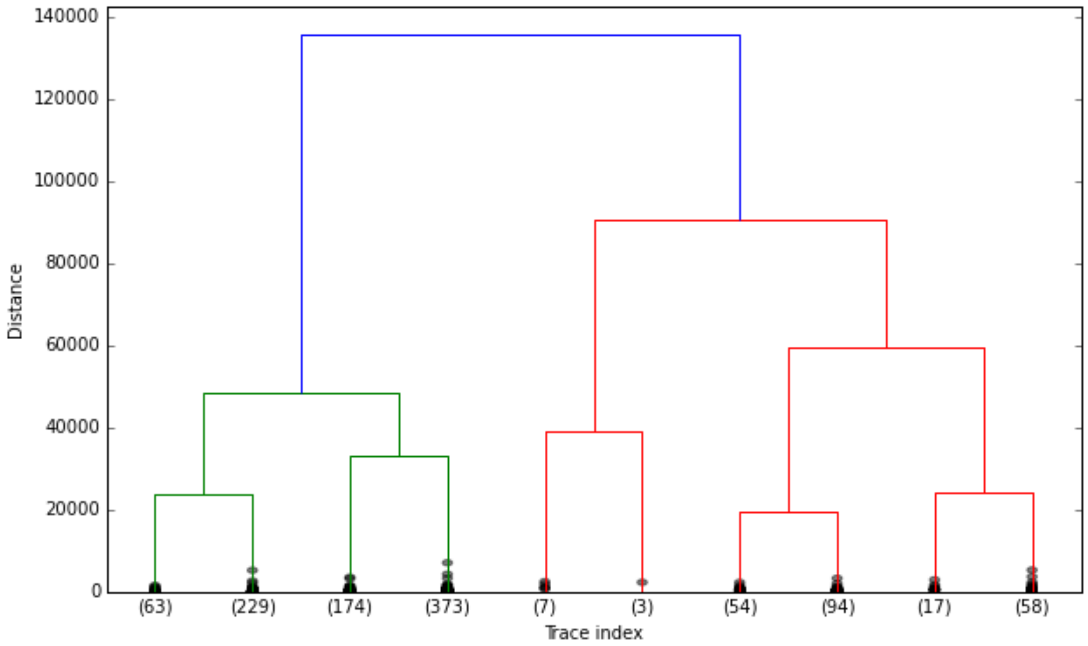


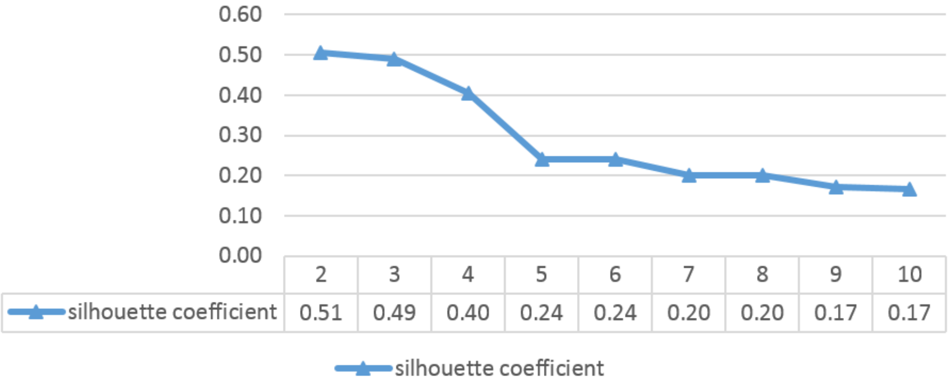
Figure 5.8 Hierarchical clustering dendogram. X axis shows trace number grouped into 10 clusters at the lowest level, Y axis shows distance based on the edit distance matrix.

Based on the hierarchical clustering result, the next step was to determine the optimal number of clusters. There are two ways to do this, which are internal and external methods. Internal methods use the information of the clustering to evaluate the goodness of the clustering structure and can be used for estimating the number of clusters without any external data. Some methods to measure internal information of the clustering are Davies-Bouldin index, Dunn index, and Silhouette coefficient. External methods compare the clustering result to externally known labels, such as the provided class labels. In this experiment, the internal methods are suitable because there was no prior knowledge of the class labels.

An internal method used in this experiment was the silhouette coefficient [176], which reflects the compactness/ cohesion (how close are the objects within the same cluster) and separation (how well-separated a cluster is from other clusters) of the cluster partitions. The silhouette coefficient measures how well an object is clustered and estimates the average distance between clusters. The formula is

$$\boldsymbol{S_i = (b_i - a_i)/\max\,(a_i, b_i)}$$

where ***b**_i* is the dissimilarity between ***i*** and its neighbour cluster and $\boldsymbol{a_i}$ is the average dissimilarity between ***i*** and all other points of the cluster where ***i*** belongs. Clustered items with a high average of silhouette value are considered well clustered. The summary of the silhouette coefficient for two to ten clusters is presented in Figure 5.9.



	2	3	4	5	6	7	8	9	10
silhouette coefficient	0.51	0.49	0.40	0.24	0.24	0.20	0.20	0.17	0.17

Figure 5.9 Silhouette coefficient of *k* = 2 to *k* = 10. The silhouette coefficient decreases as the number of clusters (k) increases.

Cluster 1 consist of traces of 839 patients (78%). The process model of cluster 1 (see Figure 5.10) is very similar to the general process model presented in Figure 5.7. The median duration is 24.4 months and the mean duration is 28.9 months. The most frequent variant consists of patients who received six cycles of *Chemotherapy (S1)* (n=21, 2.5%). The second variant consists of six *S1* followed by an *Urgent outpatient (D3)* (n=20, 2.4%). The third variant consists of six *S1* followed by a *Not neutropenic (S4)* (n=10, 1%). The average number of events per patient is 16 (min=1, max=46). Further analysis compared the trace frequency of each activity. The average difference is 5%. The highest difference is in the trace frequency of *Admission not neutropenic* (5%), where 50% of patients in the complete dataset had this activity but only 40% of patients in this cluster had this activity. The other activities had 7% or less difference on the trace frequency between this cluster and the complete dataset. The process model of cluster 1 is presented in Figure 5.10.

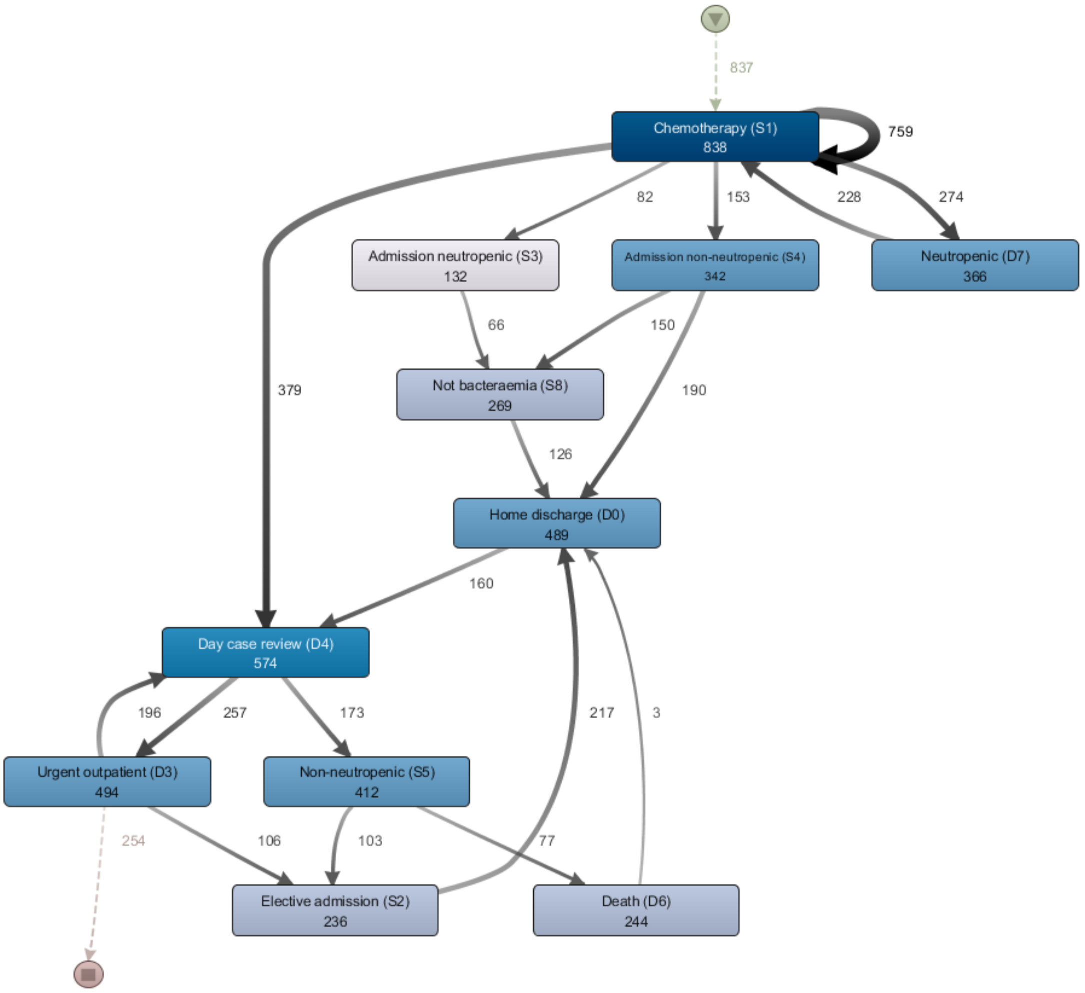


Figure 5.10 Process model of cluster 1. The process model is very similar to the general process model.

The process model of cluster 2 in Figure 5.11 consists of patients having different/ exceptional traces. The median duration is 36.4 months and the mean duration 37.8 months, both showing longer durations compared to cluster 1. Events per patient ranged from 4 to 220, with a mean of 41. There are three activities that are frequent in cluster 2 but not frequent in cluster 1 nor the complete log, which are Admission non-neutropenic (n=199/ 85%), Non-neutropenic (n=191/ 82%), and Not bacteraemia (n=156/ 66%). The process model of cluster 2 is presented in Figure 5.11.

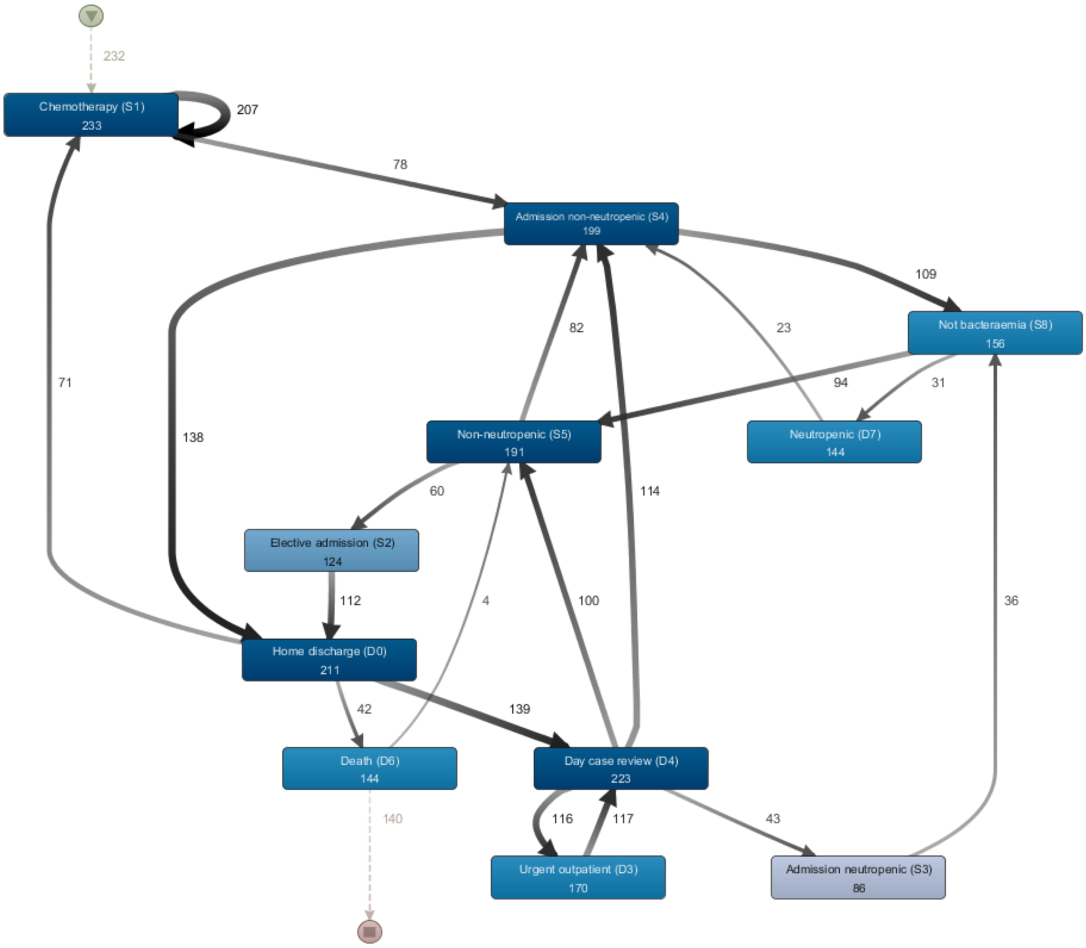


Figure 5.11 Process model of cluster 2. It shows process model of patients having exceptional care pathways (n=233/ 22%).

5.3.4 Stage 4: Evaluation

This experiment aimed to find an alternative way to group patients based on their similarities. The clustering method was separating the common traces from the exceptional ones. As presented hierarchically in Figure 5.8, two clusters were separating 838 traces (78%) from 233 traces (22%). The next level created three clusters of 839 traces (78%), 10 traces (1%), and 223 traces (21%). The final clusters were also not meaningful to separate patient groups in cancer treatment.

An insight gained from this experiment is that there is an opportunity to create a new method or to improve the available methods for analysing process variants in cancer treatment. Another insight was that the differences between clusters could be characterised by the differences of activity frequency, as described in Section 5.3.3. Trace clustering would not be explored further in this thesis, but other approaches to characterised differences of process models based on trace duration and activity frequency will be used in the next case study.

5.4 Summary

This chapter has described the analysis of the PPM Chemotherapy dataset. The PPM Chemotherapy dataset was used as an extract of the PPM database with a specific focus on records of patients receiving chemotherapy treatment within the LTHT. The analysis was initially done by reproducing the clinical pathway analysis in the previous study, with an improvement in the structured method proposed in this study. Discussions with clinical experts focused on how the visualisations resulting from the process mining approach can be useful to support understanding of the pathways.

An additional analysis of the PPM Chemotherapy dataset was done to explore the possibility of using a trace clustering approach to find different patterns within a process. The results, as presented in Section 5.3, show two clusters with visual differences compared to the complete dataset. But it appears that the trace clustering was done to separate the outlier/ the infrequent traces from the main cluster. A lesson learned from this experiment was that trace clustering can be used to find different patterns within a process, but a better approach is needed for process change analysis.

This case study can be seen from different perspectives, as described in Section 1.2. From the health service perspective, the process model in Figure 5.3, for example, are useful for the health service manager for understanding the pathway of patients receiving chemotherapy. From the process mining perspective, this case study has been shown how the dataset can be used for process mining in healthcare, despite some limitations due to the de-identification process. From the information system perspective, the conformance-based change detection in Section 5.2.3 shows that conformance values over the years are not constant. This finding implies that the process has been changed. It has been confirmed by a member of the PPM developer team that the detected changes were related to the introduction of the PPM software in 2003 and a major improvement in 2007.

The analysis of the PPM Chemotherapy dataset can be seen as an exercise to get used to the complete data in the PPM Cancer dataset. Once the ethical access to the PPM Cancer dataset was granted, this study then used the dataset in the next experiments.

Chapter 6
Case study 3: Experiments using the PPM Cancer Data

Analysis of the Patient Pathway Manager (PPM) Chemotherapy dataset produced a range of useful results that have been presented in Chapter 5. This chapter presents the analysis of the PPM Cancer dataset through four experiments. The PPM Cancer dataset also includes the patients selected in the Chemotherapy dataset.

Section 6.2 has been presented in an invited joint presentation at the Public Health England National Cancer Registration and Analysis Service (PHE NCRAS) seminar in 2018. Section 6.3 was presented and published in the 2019 Process-Oriented Data Science for Health (PODS4H) entitled "A multi-level approach for identifying process change in cancer pathways" [177]. A journal paper is prepared to extend that paper into the International Journal of Environmental Research and Public Health (IJERPH) Special Issue of PODS4H19. Section 6.4 summarises another journal paper that is prepared for a jointly-authored publication entitled "Process mining to explore variations in the 62-day pathways of endometrial cancer" [178]. Section 6.5 presents additional work with process change analysis on the whole PPM Cancer dataset.

6.1 Data description

Overview of the PPM Cancer dataset has been presented in Section 1.4.2 and has been described in Section 3.4.3. More details of this dataset are presented in this section.

6.1.1 Data provenance

The PPM Cancer dataset was generated from direct access to the PPM Query database. This database is a copy of the live database of the PPM EHR system, which contains data on patient treatment in the LTHT from 1990 until the present. The characteristics of the PPM EHR system in the LTHT were presented in Section 5.1.1.

The database contains information about all patients within the hospital, including cancer patients. Access to this database was made possible through the hospital infrastructure in the secure environment (UoL IRC). Along with the PPM Query database, other resources are also accessible. Those are the PPM Splunk, a web-based application management that captures real-time user access, and the PPM JIRA, a project management software used by the development team in the LTHT.

6.1.2 Representativeness

The PPM Query database is an exact copy of the PPM live database, as used in the LTHT. PPM is a mature EHR that captures the comprehensive clinical data of all patients receiving treatment within the LTHT. The PPM Cancer dataset consists of the clinical data of more than 3 million patients, of which more than 270,000 patients have at least one cancer-related diagnosis (the number is growing). The system integrates data from multiple systems within the LTHT, including patient admissions, treatments (chemotherapy, surgery, and radiotherapy), pathology, investigations, Multidisciplinary Team (MDT) meetings, consultations, and outpatients. The PPM dataset is one of the largest EHRs in the UK, and this makes this data highly representative for cancer treatment analyses.

6.1.3 Data characterisation

The PPM Query database was accessed based on ethics approval. It contains 49 tables in the Leeds table schema (see Appendix E.1). Every table was checked to be used in process mining to analyse patient pathways during their cancer treatment. The main step was to check each column in those 49 tables to find potential columns to be a case id, an activity name, a resource name, and a timestamp that refer to an event in patient treatment. This step needed several iterations of a careful investigation of each table, documentation review, and discussion with clinical experts.

One important challenge was that one table might contain more than one activity, and the activity names might need to be inferred from the recorded timestamp. For example, *Admissions* table has some potential columns: *PatientID, AdmissionDate, ContactSpecialityLabel, DischargeDate, DischargeMethodLabel*. There were two activities identified in this table, which were *Admission* and *Discharge* that was inferred from *AdmissionDate* and *DischargeDate,* respectively. The *PatientID* was identified as the case id. *ContactSpecialityLabel* was identified as the resource name of *Admission* and *DischargeMethod Label* was identified as the resource name of *Discharge.* The timestamps of *Admission* and *Discharge* were *AdmissionDate* and *DischargeDate*, respectively.

Overall, 12 out of 49 tables had the minimum requirements for process mining and were selected to extract patient pathways. This selection is presented in Appendix E.2. The list of selected columns from each table for process mining is shown in Table 6.1.

Table 6.1 Detail of columns for process mining

** if NULL, use cn_ContactTypeCode_CodeLabel from Contact reference table*

#	Table	case id	resource	timestamp
1	Admission	em_PatientID	em_ContactTypeLabel	em_AdmissionDate
2	ChemoCycles	ecc_PatientID	ecc_CycleContactTypeLabel	ecc_CycleStartDate
3	Consultation	eb_PatientID	eb_ContactTypeLabel	eb_ConsultationDate
4	Diagnosis	dx_PatientID	dx_ContactTypeLabel	dx_DiagnosisDate
5	Investigation	en_PatientID	en_ContactTypeLabel*	en_EventDate
6	MDT Review	ev_PatientID	ev_ContactTypeLabel*	ev_EventDate
7	Outpatient	op_PatientID	op_ActualContactTypeLabel	op_ClinicDate
8	Pathology	esp_PatientID	esp_ContactTypeLabel*	esp_PathologyDate
9	Patients	pt_PatientID	NULL	pt_DeathDate
10	Radiotherapy	er_PatientID	er_ContactTypeLabel	er_EventDate
11	Referral	ef_PatientID	ef_SourceCodeLabel	ef_ReferralDecisionDate
12	Surgery	es_PatientID	es_ContactTypeLabel*	es_SurgeryDate

6.1.4 Data quality

The data quality of the PPM Cancer dataset was assessed based on the data quality framework for process mining of EHR data [179]. This framework was built based on methods and dimensions of data quality assessment [133]. An important approach proposed in this framework was to register potential data quality dimensions and analyse them. The result is summarised in the following paragraphs.

The *completeness* of the data was assessed with element presence, data element agreement, data source agreement, distribution comparison, and validity checking methods. The main concern was to find missing data in case, event, activity name, and timestamp. Completeness checking on the 49 tables in the PPM Cancer dataset was done by analysing the data and discussing the results with the clinical experts. One important finding was that the completeness of the data changes over time. This is because the PPM system is evolving. For example, *Outpatient* activity started to appear in patient records in 2006. This is not because patients before 2006 were never seen in any *Outpatient* activities, but because this activity just started to be recorded in the PPM system in 2006.

The *correctness* of the data was assessed with element presence, data source agreement, and validity checking methods. The main focus was to analyse imprecise data in case, event, activity name, and timestamp. One important finding is that the *Surgery* table contains event data of all types of surgeries, including diagnostic surgery and therapeutic surgery. A clinical expert categorised the main procedures of the surgeries to separate those two. As a result, there are two activities derived from the surgery events: *Diagnostic Surgery* and *Surgery*. Another issue found during the analysis is that some events were recorded with the date and 00:00:00 was added as the hour, minute and second. This led to a problem where those events seem to have occurred in the middle of the night, when they actually did not.

The *concordance* of the data was assessed with element presence, data source agreement, and distribution comparison methods. The main focus was to analyse irrelevant data in case, event, activity name, and timestamp. This was done by examining columns in all tables within the PPM database to find relevant columns for process mining. There are 12 tables relevant for process mining of cancer treatment pathways, and some other tables used as reference tables. For example, the *Contacts* table was used as a reference table for defining the resource from the original table.

The *plausibility* of the data was assessed with element presence and distribution comparison. The main focus was to find incorrect data in case, event, activity name, and timestamp. One example of an issue found in the plausibility of the PPM Cancer data was the surgery types. Early iterations on the pathway analysis found that the number of surgery records was too high, suggesting that the majority of patients had undergone surgery in their cancer treatment. Further discussions with clinical experts found that this was because surgery can be categorised into diagnostic surgery and [therapeutic] surgery. The issue of 00:00:00 as default time is also a critical problem in process mining. When the sequence of events is analysed, the events with 00:00:00 time would be treated as the first event during the day and obstruct the real sequence of events.

The *currency* of the data was assessed with a log review. The assessment was to check if they were recorded in the PPM EHR within a reasonable period of time following the activity. One issue found in the currency of PPM Cancer dataset was that there are records dated back to the 1990s, while the PPM EHR had only begun to be used in the LTHT in 2003. Discussions with the development team suggested that these data were manually recorded in the PPM database by the team in the early years of the PPM EHR system use. Another possible issue was data error, which might have occurred during data recording. To handle this issue, the analysis did not use any data before 2003.

The most crucial issue in the process mining of the PPM Cancer dataset is that the PPM database is supporting PPM software, an EHR system developed in the LTHT. The system was inevitably changed over time for many reasons. Some examples of the reasons are the changes needed when a clinician found an error in the system, the changing guidelines for the treatment of a specific type of disease, and the decision taken by the hospital to connect with other services outside the hospital. The naturally changing environment made it challenging to analyse the data with process mining. The process mining results for a subset of the data during a specific period might be different from those from another period.

6.1.5 Data variety

In the complete set of the PPM database, data variety is high. The PPM database was focused on recording data in cancer treatment since 2003 and has been underpinned to manage all data within the whole Trust since 2012. Further improvement was made in 2014 to join the Leeds Care Record (LCR), an integrated digital care record system across the Leeds city region. Those major changes are important as starting points to help understand the high complexity and high variety of the PPM dataset. For this research, data analysis focused on cancer patient records in the PPM database.

The variety of cancer patients can be identified by the number of patients diagnosed with cancer during 2002–2017, as presented in Table 6.2. The most common cancer type in the PPM Cancer set is the C00–C75 (61.65%), followed by C81–C96 (55.55%) and D10–D36 (26.35%). One patient might have more than one type of cancer and can be included in more than one group.

Table 6.2 The number of patients based on cancer types

ICD code	Cancer type	Patients	%
C00-C75	Malignant neoplasms, stated or presumed to be primary, of specified sites, except of lymphoid, haematopoietic and related tissue	236,244	61.65
C76-C80	Malignant neoplasms of ill-defined, secondary and unspecified sites	7,203	1.88
C81-C96	Malignant neoplasms, stated or presumed to be primary, of lymphoid, haematopoietic and related tissue	21,285	55.55
D00-D09	In situ neoplasms	12,546	3.27
D10-D36	Benign neoplasms	100,989	26.35
D37-D48	Neoplasms of uncertain or unknown behaviour	4,930	1.29
Total		**383,197**	**100.00**

6.1.6 Limitations of using PPM Cancer dataset in this study

The limitations of using the PPM Cancer dataset in this research are closely related to the challenges of analysing real-life datasets. The PPM EHR system is a growing system, with a development team working continuously to improve and make changes as required by the hospital. This condition causes the quality of the PPM Cancer dataset varies over time. The long duration of data stored in the PPM Cancer dataset means it is not possible to assume that no changes happened to the process over that time. It is also not possible to understand the complete history of the changes that happened to the system based on the documentation only.

One approach for change analysis is to separate one change and investigate the effects on the related process. Another approach is to analyse the process and detect the changes based on a specific pattern of interest (trend, seasonal, or residual patterns).

6.2 Experiment 5: GP tab change analysis

As mentioned in Section 6.1.6, the most critical issue in using the PPM database is that the system was changed over time. This study focused on the changes evidenced in the User Interface (UI) of the PPM EHR system, but the reasons for the UI changes are out of this discussion. This limitation was because the reason behind a change might not be evidenced in the data.

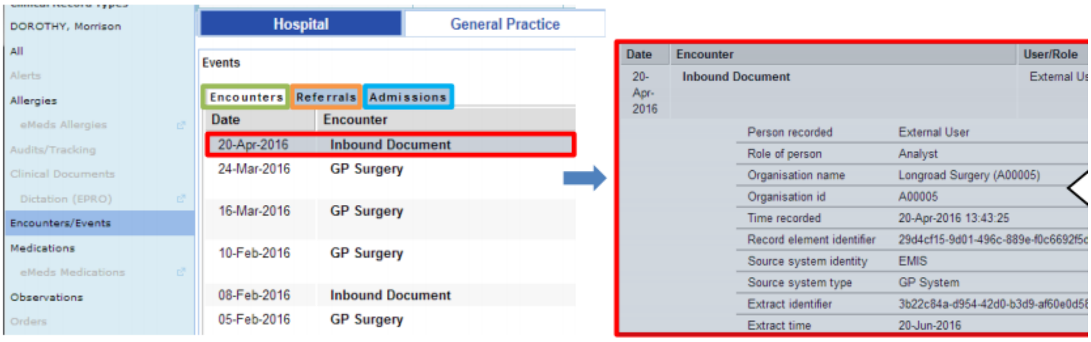
All user interactions are recorded in the PPM Splunk, and every time a user views data in the PPM EHR system, the system automatically updates the data in the system. Whenever the PPM EHR system is changed, those changes are recorded in the PPM JIRA software. The PPM developer team uses the PPM JIRA software as a dashboard to track issues and changes to the PPM EHR system. From the PPM JIRA, detailed information about the change can be found in timely order, including the request for change, processing, and the release of the new updated version.

In this research, UI changes in the PPM Cancer dataset were explored in two types of experiments. The first one was to analyse a process change when a change is known. For this purpose, the introduction of a General Practitioner (GP) Tab was chosen as a change of interest. The experiment using the GP Tab change is presented in this section. The second type of experiment was to detect changes without prior knowledge about any change and is presented in the next sections.

6.2.1 Stage 1: Planning and justification

This stage was done to plan and justify the experiment. This experiment aims to examine the effects of GP tab introduction in cancer treatment. The primary research question is “*Is it possible to analyse process changes from a given UI change?*”. The GP Tab introduction is one example of a UI change in the PPM EHR system. The GP Tab allows clinicians to access patient records within the GP system required to support clinical decisions for patient treatment. This requires a change in the PPM EHR system, to provide access to the GP records.

The GP Tab is a feature in the PPM EHR system presenting the GP information (such as diagnosis, allergies, and medications) recorded for patients registered with a Leeds GP taking part in the LCR. This GP Tab was introduced in July 2014. An illustration of the GP Tab in the PPM EHR system is presented in Figure 6.1.



Date	Encounter
20-Apr-2016	Inbound Document
24-Mar-2016	GP Surgery
16-Mar-2016	GP Surgery
10-Feb-2016	GP Surgery
08-Feb-2016	Inbound Document
05-Feb-2016	GP Surgery

Date	Encounter	User/Role
20-Apr-2016	Inbound Document	External Us

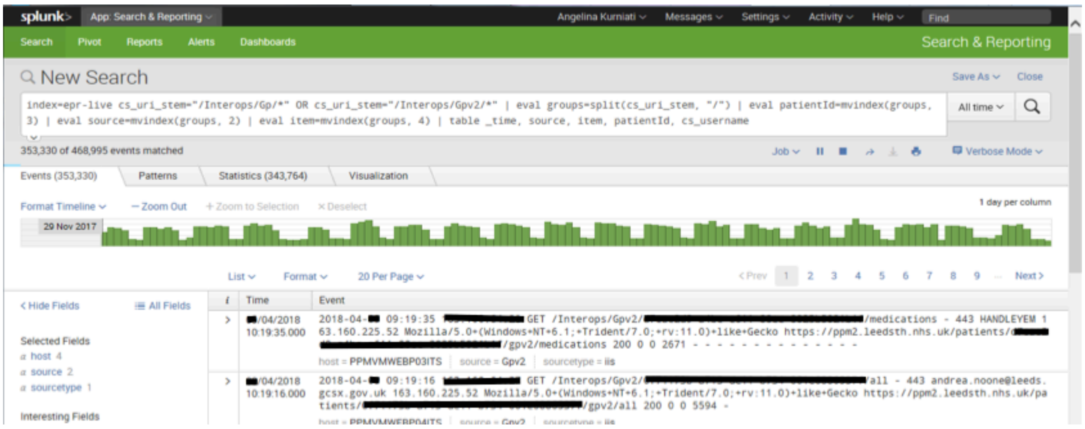
Person recorded	External User
Role of person	Analyst
Organisation name	Longroad Surgery (A00005)
Organisation id	A00005
Time recorded	20-Apr-2016 13:43:25
Record element identifier	29d4cf15-9d01-496c-889e-f0c6692f5c
Source system identity	EMIS
Source system type	GP System
Extract identifier	3b22c84a-d954-42d0-b3d9-af60e0d58
Extract time	20-Jun-2016

Figure 6.1 The GP tab in the PPM EHR system. This preview is from the official website of PPM support [180].

In this experiment, the usage of the GP Tab in the PPM EHR system is recorded in the PPM Splunk. The record of GP Tab usage was then analysed to explore the pattern of this usage over time. The record was also transformed into an event log and combined with the event log of the cancer treatment. The combined event log was used to explore the effect of the GP Tab access on the cancer treatment. For this experiment, the scope was breast cancer patients receiving chemotherapy from 2014 to 2018. This experiment included clinical experts in every stage.

6.2.2 Stage 2: Extraction, transformation, and loading

The extraction was done through a query in the PPM Splunk. The view of the query page in PPM Splunk is presented in Figure 6.2. This view contains detailed data on the *date* and *time*, *page address*, *patient id* and *user id* recording a time when a clinician had accessed the GP Tab page of a patient. There is also a bar chart visualising the number of records on a daily basis. The bar chart shows an obvious pattern of weekday- and weekend- usages.



```
index=epr-live cs_uri_stem="/Interops/Gp/*" OR cs_uri_stem="/Interops/Gpv2/*" | eval groups=split(cs_uri_stem, "/") | eval patientId=mvindex(groups, 3) | eval source=mvindex(groups, 2) | eval item=mvindex(groups, 4) | table _time, source, item, patientId, cs_username
```

353,330 of 468,995 events matched

```
> [redacted]/04/2018   2018-04-[redacted] 09:19:35 [redacted] GET /Interops/Gpv2/[redacted]/medications - 443 HANDLEYEM 1
  10:19:35.000         63.160.225.52 Mozilla/5.0+(Windows+NT+6.1;+Trident/7.0;+rv:11.0)+like+Gecko https://ppm2.leedsth.nhs.uk/patients/[redacted]
                       [redacted]/gpv2/medications 200 0 0 2671 - - - - - - - - - - - - - - -
                       host = PPMVMWEBP03ITS | source = Gpv2 | sourcetype = iis
> [redacted]/04/2018   2018-04-[redacted] 09:19:16 [redacted] GET /Interops/Gpv2/[redacted]/all - 443 andrea.noone@leeds.
  10:19:16.000         gcsx.gov.uk 163.160.225.52 Mozilla/5.0+(Windows+NT+6.1;+Trident/7.0;+rv:11.0)+like+Gecko https://ppm2.leedsth.nhs.uk/pa
                       tients/[redacted]/gpv2/all 200 0 0 5594 -
                       host = PPMVMWEBP04ITS | source = Gpv2 | sourcetype = iis
```

Figure 6.2 Records of user access in the PPM Splunk. Confidential information such as patient ID and dates are blocked in black.

The result of the query was plotted in a bar chart representing the number of clicks on the GP Tab every day from July 2014 until December 2018. The plot is presented in Figure 6.3.

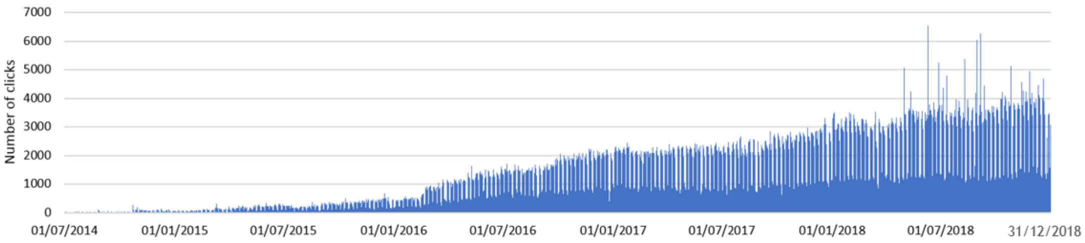


Figure 6.3 GP tab clicks each day. It shows that the number of clicks generally increased over time, with steady fluctuations showing the pattern of weekday- and weekend- usages.

In March 2018, there was an improvement of the Medical Interoperability Gateway (MIG) from MIG1 (GPv1) to MIG2 (GPv2). MIG is the gateway to integrate the GP SystmOne/EMIS into the LCR. The transition period from the GPv1 to GPv2 can be captured in the monthly usage from 2017 to 2018, as shown in Figure 6.4.

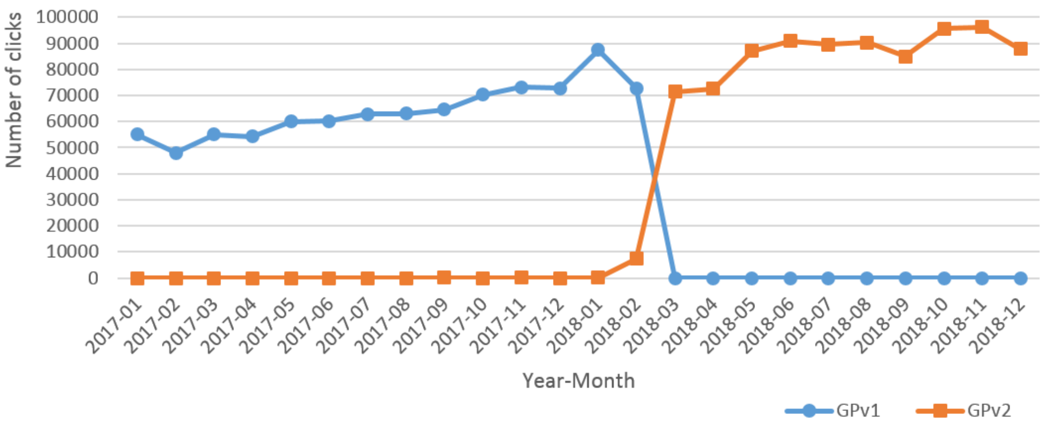


Figure 6.4 Monthly usage of GP tab during 2017-2018. The blue dots are monthly usage of the older version (GPv1) and the orange dots are those of the new version (GPv2).

During September 2017 to February 2018, both versions were accessed by the clinicians. A small number of clinicians (7 to 14 users) clicked through the GPv2 from September 2017 to January 2018. The GPv2 was later tested by a larger number of clinicians (1,117 users) in February 2018 and fully replaced the GPv1 from March 2018. The GPv1 was deprecated completely in March 2018.

6.2.3 Stage 3: Mining and analysis

Stage 3 of this experiment was undertaken to combine cancer patient events in the PPM Cancer dataset of the GP Tab access within the PPM Splunk. The initial analysis was done by checking the intersection of cancer patients in the PPM Splunk and the PPM Cancer records. The GP Tab access in the PPM Sclunk is from July 2014 to April 2018. The cancer patients included in the experiment were those diagnosed with C% or D% during 2002–2017. There were 46,547 out of 339,127 cancer patients (37%) who had their GP Tab clicked by clinicians. On the other hand, there are 46,547 out of 171,468 GP Tab access (16%) are the GP Tab of cancer patients.

Process mining was done to continue the analysis of the chemotherapy treatment of breast cancer patients, as presented in Section 5.2. This experiment analyses Leeds patients diagnosed with breast cancer (C50) who received EC-90 as adjuvant chemotherapy from 2014 to 2018, and whose GP Tab was clicked by clinicians. There were 733 patients included in this selection. The analysis was done to explore GP Tab access during chemotherapy cycles. The event log was a combination of patient events in the PPM Cancer dataset and the GP Tab access of those patients in the PPM Splunk.

Figure 6.5 shows the process map containing the flow from Cycle 1 of chemotherapy to the subsequent cycles up to Cycle 6. During the course of chemotherapy, the GP Tab might be accessed by clinicians. The most frequent sequence is that the GP Tab was accessed after Cycle 6 for 160 out of 339 GP Tab clicks (47%). This was discussed with clinical experts. The clinicians might need to check on patient records in the GP Tab after the sixth cycle to decide whether to discharge the patient or to suggest another treatment. The next most frequent one was after Cycle 3 for 110 clicks (32%). The possible case was that clinicians needed to check on patient records in the GP Tab after Cycle 3, to decide if the next cycles should be delivered as planned or not. Another interesting finding was that in 326 clicks (96%), the GP Tab click is the last activity in the pathway, or at the end of the treatment.

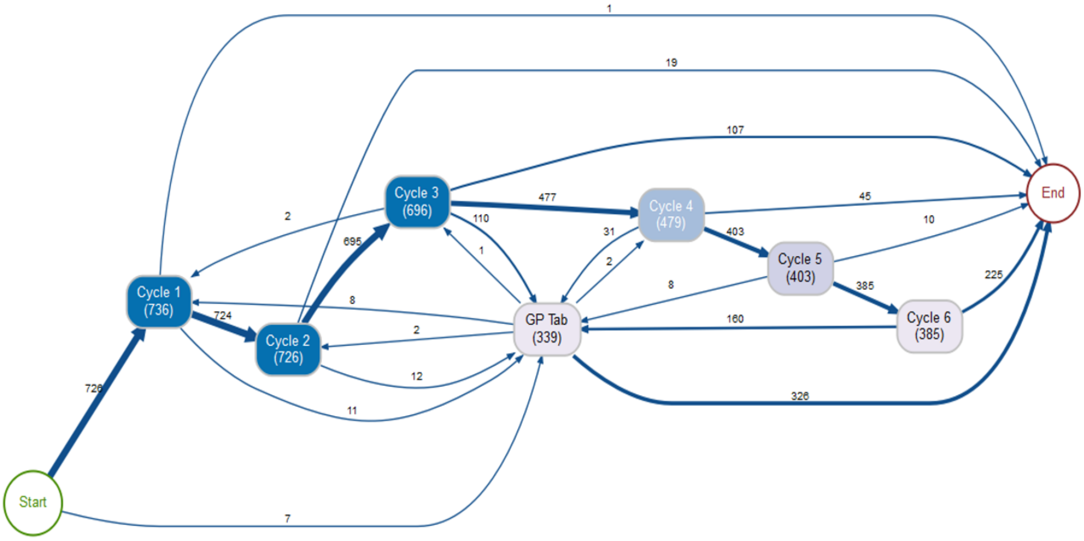


Figure 6.5 Process model of GP tab access during chemotherapy cycles. This was built using bupaR. It shows that GP Tab was mostly accessed after *Cycle 3*, *Cycle 6*, or *Cycle 4*.

6.2.4 Stage 4: Evaluation

In this experiment, the evaluation was done in both statistical and clinical aspects. The results of the statistical evaluation were presented along with each result in Stage 3 above. The clinical evaluation was done through discussion with clinical experts and is presented in the following paragraphs.

In *Stage 1*, clinical experts suggested the scope of the study. Some known changes in the PPM EHR system were discussed with the clinical experts, the development team, and the software training team. The GP Tab change was chosen based on the availability of the related data to explore process change. One important insight from the software training team was that for some new features introduced in the PPM software, there was a period when training was given to the clinicians to introduce the use of a new feature. This training period might affect the analysis of system usage. It was found that the training and testing period of the GP Tab introduction was from September 2017 to February 2018.

In *Stage 2*, clinical experts evaluated the extraction step and suggested some changes. This included a change to focus on the effect of the GP tab introduction to the chemotherapy cycles. The decision to focus on one specific process and one change of UI was chosen to localise the change and the effects of that change to the treatment process. It was followed by a discussion with the programme manager of Leeds Care Record (LCR). LCR is the initiative to create integrated records of patients across providers and between different systems in Leeds. The GP tab introduction was part of the LCR program. The approach undertaken in this experiment was said to be promising to analyse the effect of the introduction of different functionality in the LCR program. The additional data from the PPM JIRA was also suggested by the development team to get complete records of the change. The PPM JIRA was useful to gather as many information as possible about a change being analysed.

In *Stage 3*, the evaluation was done through a discussion of the results from the clinical perspective. The idea of combining user access records in PPM Splunk with the treatment records in the PPM Query database was good to analyse the effect of a system change to the treatment process. Another possibility discussed was to analyse PPM Splunk separately to be compared to the process model discovered from the patient record. Since PPM Splunk recorded all actions done by clinicians during patient treatment, the treatment process itself should be reflected in the records.

6.3 Experiment 6: Endometrial cancer pathways from referral to diagnosis

This experiment was done to explore process change over time without a prior known change. The endometrial cancer pathway was chosen in this experiment because of clinical expert availability. This was also because endometrial cancer is one of the most common cancers in the gynaecology department, and it was understood that the procedure for endometrial cancer had not been changed radically within the last 15 years. In this study, analysis of the endometrial cancer pathway was focused on the change process analysis, to detect and analyse change points over time. Based on a discussion with clinical experts, analysis can be done into two steps. The first step is the pathway from GP referral to diagnosis and is presented in this section. The second step is the pathway from GP referral to first treatment that represents the 62-day pathway and is presented in Section 6.4.

This section explores the process change analysis of endometrial cancer treatment. This experiment followed the general methodology described in Chapter 3. The changes found were analysed to find process evolution within the system. The complete stages have been published [177] and the important findings are summarised in Sections 6.3.1 to 6.3.4.

6.3.1 Stage 1: Planning and justification

Stage 1 (planning and justification) in this experiment was done by understanding the data and additional resources of PPM development [21]. The scope of the experiment is the analysis of Leeds patients diagnosed with endometrial cancer during 2003–2017. The analysis focused on the pathway from GP referral to the diagnosis of endometrial cancer. The research questions are:

1) *What is the pathway of endometrial cancer from referral to diagnosis?*
2) *Are there differences in care paths over time?*

The experiment involved clinical experts and a statistician during all stages of the study. This was done through regular discussions at the end of each stage.

6.3.2 Stage 2: Extraction, transformation, and loading

The ***extraction*** was done by obtaining all the events that happened to the selected cohort from GP referral to diagnosis. The criteria were to include: (1) all patients who have a legitimate care relationship with the LTHT, (2) those who have a definitive/primary diagnosis of endometrial cancer (C54 and C55), (3) those with a GP referral as the start event and the diagnosis as the end event, and (4) those having a maximum duration of 120 days. These selection criteria were applied in a database query to create an event log. The extraction was done by selecting columns for process mining: *case_id, activity, resource,* and *timestamp*. For this experiment, there were 943 patients with a total of 96,067 events in the event log.

The ***transformation*** was done to filter the extracted event log as required. This included removing missing values, merging subsequent events, and adding artificial START and END events. Two additional filtering processes were performed to split *Surgery* and to handle the same-day events. *Surgery* was split based on two types of surgery (diagnostic and therapeutic surgeries) into *Diagnostic Surgery* and *Surgery* activities. Same-day events were ordered based on the logical sequence of the events as discussed with the clinical experts. The number of events and patients having those events for each activity in the final event log is presented in Table 6.3.

Table 6.3 Detail of events in the event log

#	Activity	Events	n	n (%)
1	*Admission*	605	519	55%
2	*Chemotherapy*	1	1	0.1%
3	*Consultation*	357	152	16%
4	*Death*	1	1	0.1%
5	*Diagnosis*	943	943	100%
6	*Diagnostic Surgery*	972	809	86%
7	*Discharge*	284	256	27%
8	*Investigation*	1,759	738	78%
9	*MDT Review*	395	231	24%
10	*Outpatient*	228	149	16%
11	*Pathology*	1,239	860	91%
12	*Radiotherapy*	1	1	0.1%
13	*Referral*	1,854	943	100%
14	*Surgery*	274	258	27%
Total		**8,915**	**943**	**100%**

* *n = number of patients*

Table 6.3 shows that all patients have *Referral* and *Diagnosis*, as these are required as the start and end events of the process. The most infrequent activities are *Chemotherapy, Radiotherapy*, and *Death*. Along with *Surgery*, *Chemotherapy* and *Radiotherapy* are the treatment types for cancer. Those treatments commonly happen after diagnosis. In some cases, the diagnosis and the first treatment happened on the same day, hence they are extracted from this case study.

6.3.3 Stage 3: Mining and analysis

The mining and analysis stage in this study was done through process mining and process analytics. Process mining includes process discovery and conformance checking. Process analytics was performed to analyse process change over time using a multi-level process comparison approach and is presented in Section 6.3.4.

Process discovery was performed using iDHM, as a plugin in ProM that provides many options for process model abstraction. The input is an event log created as presented in Section 6.3.2. The directly-follows graph is shown in Figure 6.6 below.

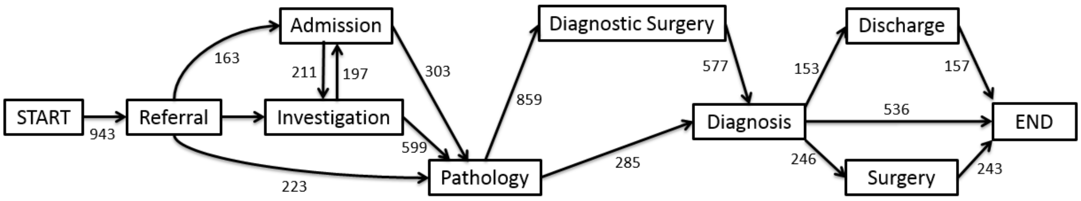


Figure 6.6 The directly-follows graph, showing process model of the pathway.

Figure 6.6 shows the most frequent pathways from referral to diagnosis of endometrial cancer in the 943 patients. For simplicity, the process model shows the eight most frequent activities and the most frequent paths between them. The *Outpatient*, *Consultation*, and *MDT Review* activities were omitted to produce a simple diagram. The *Outpatient* activity appeared in 149 out of 943 patients (16%), *Consultation* appeared in 152 patients (16%), and *MDT Review* appeared in 231 patients (24%).

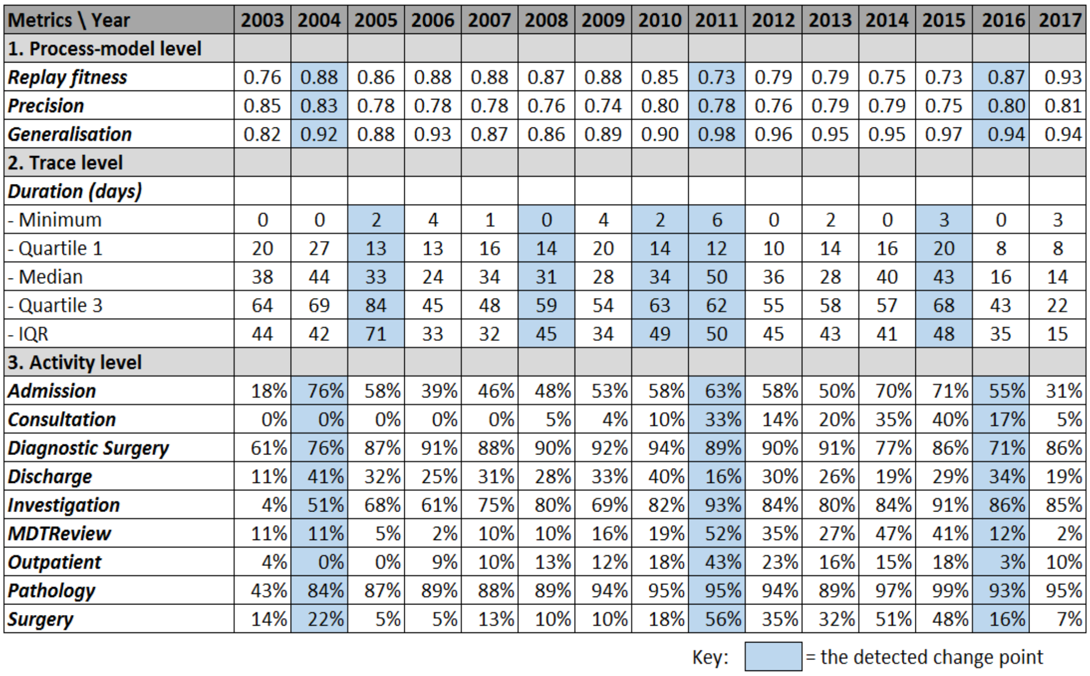
Conformance checking was performed to check conformance of the reference model to the traces in the event log [86]. The reference model was discovered from the complete event log using iDHM plugin. The general process model was highly representative of the complete event log with a replay fitness of 0.809, a precision of 0.83, and a generalisation of 0.996. This means that the process model can accurately reproduce the traces recorded in the log, allows behaviours that are not seen in the event log, and can reproduce predicted future behaviours of the process.

6.3.4 Process change analysis with a multi-level approach

The analysis of process change over time was done by comparing the process in three levels of detail: process model, trace, and activity levels, as presented in Table 3.2 in Section 3.3.1. The idea was to explore process changes by comparing process execution over time within those three levels. The process-model level compares replay fitness, precision, and generalisation of the general process model to the traces over time. The trace level compares duration and variant proportion over time. The variant proportion compares the proportion of the general variants in the sub-logs. The activity level compares the frequency and percentage of the activities over time.

A summary of the ***process-model comparison*** results is presented in Table 6.4. It shows that the general process model was representative of each year log.

Table 6.4 Summary of multilevel comparisons



Metrics \ Year	2003	2004	2005	2006	2007	2008	2009	2010	2011	2012	2013	2014	2015	2016	2017
1. Process-model level															
Replay fitness	0.76	0.88	0.86	0.88	0.88	0.87	0.88	0.85	0.73	0.79	0.79	0.75	0.73	0.87	0.93
Precision	0.85	0.83	0.78	0.78	0.78	0.76	0.74	0.80	0.78	0.76	0.79	0.79	0.75	0.80	0.81
Generalisation	0.82	0.92	0.88	0.93	0.87	0.86	0.89	0.90	0.98	0.96	0.95	0.95	0.97	0.94	0.94
2. Trace level															
Duration (days)															
- Minimum	0	0	2	4	1	0	4	2	6	0	2	0	3	0	3
- Quartile 1	20	27	13	13	16	14	20	14	12	10	14	16	20	8	8
- Median	38	44	33	24	34	31	28	34	50	36	28	40	43	16	14
- Quartile 3	64	69	84	45	48	59	54	63	62	55	58	57	68	43	22
- IQR	44	42	71	33	32	45	34	49	50	45	43	41	48	35	15
3. Activity level															
Admission	18%	76%	58%	39%	46%	48%	53%	58%	63%	58%	50%	70%	71%	55%	31%
Consultation	0%	0%	0%	0%	0%	5%	4%	10%	33%	14%	20%	35%	40%	17%	5%
Diagnostic Surgery	61%	76%	87%	91%	88%	90%	92%	94%	89%	90%	91%	77%	86%	71%	86%
Discharge	11%	41%	32%	25%	31%	28%	33%	40%	16%	30%	26%	19%	29%	34%	19%
Investigation	4%	51%	68%	61%	75%	80%	69%	82%	93%	84%	80%	84%	91%	86%	85%
MDTReview	11%	11%	5%	2%	10%	10%	16%	19%	52%	35%	27%	47%	41%	12%	2%
Outpatient	4%	0%	0%	9%	10%	13%	12%	18%	43%	23%	16%	15%	18%	3%	10%
Pathology	43%	84%	87%	89%	88%	89%	94%	95%	95%	94%	89%	97%	99%	93%	95%
Surgery	14%	22%	5%	5%	13%	10%	10%	18%	56%	35%	32%	51%	48%	16%	7%

Key: ▨ = the detected change point

The median [Interquartile Range (IQR)] of the replay fitness is 0.86 [0.10], the precision is 0.78 [0.03], and the generalisation is 0.93 [0.06]. All three conformance measures were similar across all years. The exceptions were in 2004 when the precision dropped, 2011 when the replay fitness dropped and the generalisation increased, and 2016 when both trace fitness and precision started to increase. Therefore, the three potentially significant change points are 2004, 2011, and 2016.

The ***trace comparison*** was done by examining the trace duration and variant proportion for each yearly sub-log. There is no obvious qualitative pattern in the trace duration, except on the IQR. The IQR generally decreases across the years, with the exception of the increasing IQR on 2005 from 42 to 71 days (68%), on 2008 from 32 to 45 (39%), on 2010 from 34 to 49 days (44%), on 2011 from 49 to 50 days (2%), and on 2015 from 41 to 48 days (18%). Based on this analysis, five periods were detected as potential changes: 2005, 2008, 2010, 2011, and 2015. There is no obvious pattern in the variant proportion over time, except the waving trend of the first variant (*Referral* → *Investigation* → *Pathology* → *Diagnostic Surgery* → *Diagnosis*) and the decreasing trend of the other variants.

The ***activity comparison*** was done by analysing the percentage of each activity for the number of patients each year. The activities were grouped into frequent activities (≥ 60%), infrequent activities (< 60%), and high-varied activities in between. Qualitatively, the periods of 2004, 2011, and 2016 were marked by changes in the frequency of the activities. In 2004, all activities had a significant increase, except for the infrequent activities. In 2011, there are significant increases in the four infrequent activities, while *Discharge* decreased to be lower than the four infrequent activities. In 2016, the frequency of the infrequent activities was increased except for *Outpatient*.

6.3.5 Stage 4: Evaluation

In this experiment, *statistical evaluation* was presented along with the results described in Stage 3. The multilevel approach for process change analysis was found to be useful to support discussions with the statistical and clinical experts in this study.

The *clinical evaluation* was done through discussions with the clinical experts and a member of the development team within the hospital. The discovered process model reflected the general pathway from referral to diagnosis of endometrial cancer. There was no significant change in the duration and sequence of the pathways that the clinical experts were aware of, which confirmed the trace-level comparison. A concern was raised about some trace variants having an *Admission* without a *Discharge*. This is the case where the *Discharge* happened after *Diagnosis* and was not included in this experiment. One important discussion in the activity-level comparison was that the EHR system is continually evolving. Some activities had only begun to be recorded in later years, such as *Outpatient* (2006) and *Consultation* (2008). The multilevel comparison was successfully captured the system changes.

6.4 Experiment 7: The 62-day pathways of endometrial cancer

This section describes the analysis of the endometrial cancer pathway from GP referral to the first treatment. This reflects the 62-day wait pathway and is a continuation of the analysis in Section 6.3. The stages of the analysis followed the general methodology presented in Chapter 3 and are very similar to the stages of the analysis in Section 6.3. The specific steps for this experiment of the 62-day pathway are presented in this section. Those steps that are similar are not presented again in order to avoid duplication.

6.4.1 Stage 1: Planning and justification

This study expands the first experiment to answer two research questions:

1) *What is the general pathway of endometrial cancer from GP referral to the first treatment?*
2) *Are there differences in care paths followed by different patient groups?*

The first research question was addressed by analysing all the events that happened to the patients during the period from GP referral to the first treatment. The second research question was explored by creating patient groups based on the types of first treatment they received for their cancer, their age at diagnosis, and the year of diagnosis. The three types of cancer treatment are surgery, chemotherapy, and radiotherapy. The age of patients was grouped into ten-year ranges. The year of diagnosis ranges from 2003 to 2018. These grouping criteria were selected based on the clinical expert suggestion that process change over time might be related to the treatment types and age range.

6.4.2 Stage 2: Extraction, transformation, and loading

Extraction was done from the PPM Query database in the Microsoft SQL Server database management system. Patient records were extracted using R codes with embedded SQL queries. The extracted records were then transformed into an event log containing *case_id, event names, resource names,* and *timestamps.* Patient data were carefully selected from 12 tables as identified in Section 6.1.3. Event names were derived directly from the table names with further adjustment based on discussions with clinical experts. There are timestamps (down to the day) in each of those 12 tables that are used directly in this study.

The selection criteria were similar to those in the experiment in Section 6.3. The criteria were to include: (1) all patients who have a legitimate care relationship with the LTHT, (2) those who have a definitive/primary diagnosis of endometrial cancer (C54 and C55), (3) those with a GP referral as the start event and first treatment (chemotherapy, radiotherapy, or surgery as the end event, and (4) those having a maximum duration of 240 days.

The transformation was done to create an event log and adjust it to be used in process mining. The event log was filtered following the same approaches as mentioned in Section 6.3.2. There were 949 patients selected with a total of 17,413 events. The selected events were between July 2001 and May 2018. The duration of the traces ranged from 5 days to 238 days, with a median duration of 62 days and a mean of 79 days. There were 921 trace variants out of those 949 patients, suggesting a high variability of the traces among patients. In this study, the variability was explored based on the first treatment, patient age at diagnosis, and the year of diagnosis. The proportion of patients in each group is presented in Table 6.5.

Table 6.5 Proportion of patients in each group

First treatment	
Group	**N (%)**
Surgery	877 (92%)
Radiotherapy	49 (5%)
Chemotherapy	24 (3%)

Age at diagnosis	
Group	**N (%)**
20s (20-29)	3 (0.3%)
30s (30-39)	10 (1%)
40s (40-49)	54 (6%)
50s (50-59)	205 (22%)
60s (60-69)	318 (34%)
70s (70-79)	245 (26%)
80s (80-89)	105 (11%)
90s (90-99)	9 (0.9%)

Year of diagnosis	
Group	**N (%)**
2003	29 (3%)
2004	38 (4%)
2005	44 (5%)
2006	51 (5%)
2007	54 (6%)
2008	57 (6%)
2009	48 (5%)
2010	64 (7%)
2011	94 (10%)
2012	94 (10%)
2013	105 (11%)
2014	87 (9%)
2015	80 (8%)
2016	52 (5%)
2017	52 (5%)

Table 6.5 shows the high variability of the cohort, based on the three grouping criteria. The first treatment groups show that the most frequent first treatment is surgery (92%), as confirmed by the clinical experts. The most frequent age groups at diagnosis are the 50s (22%), 60s (34%), and 70s (26%). There was an increasing trend of the number of patients over the years except in years 2009 and 2014–2017. Some groups had a small number of patients in them, for example patients who were diagnosed in their 20s. This has been confirmed with clinical experts to reflect the real proportion of endometrial cancer patients. The groups with a small number of patients were not analysed further to comply with the information governance and prevent re-identification of the patients.

6.4.3 Stage 3: Mining and analysis

The ***process mining*** step was done through process discovery and conformance checking. The process model is presented in Figure 6.7.

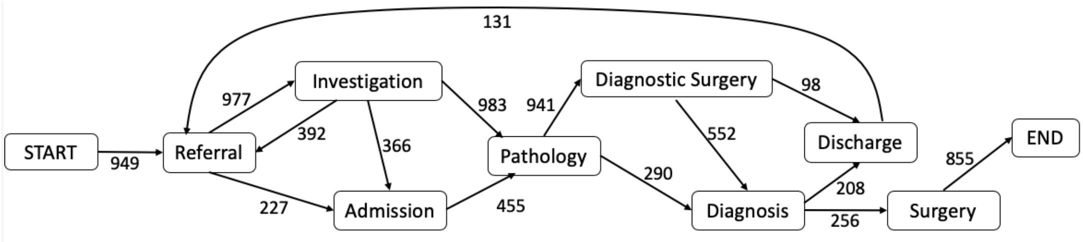


Figure 6.7 Process model of endometrial cancer 62-day pathways

Figure 6.7 shows the simplified pathways with a minimum frequency of 0.1. Conformance checking was done in ProM and this showed conformance values as follows: a trace fitness score of 0.83, a precision score of 0.80, and a generalisation score of 0.995. These values show that the model is highly representative of the general pathway of endometrial cancer treatment. The three infrequent activities not presented in the process model are *Chemotherapy* (24 or 0.1%), *Radiotherapy* (49 or 0.2%) and *Death* (1 or 0.006%). Other highly-frequent activities which are not presented in the process model, so might happen in many points within the sequence, are *Consultation, MDT Review,* and *Outpatient.*

The variability can be presented as a dotted chart, or as the distribution of the pathway duration, as shown in Figure 6.8. This shows that the highest frequency is at 62 days. A long tail of duration > 62 days reflects the real circumstances of patients with more complicated conditions.

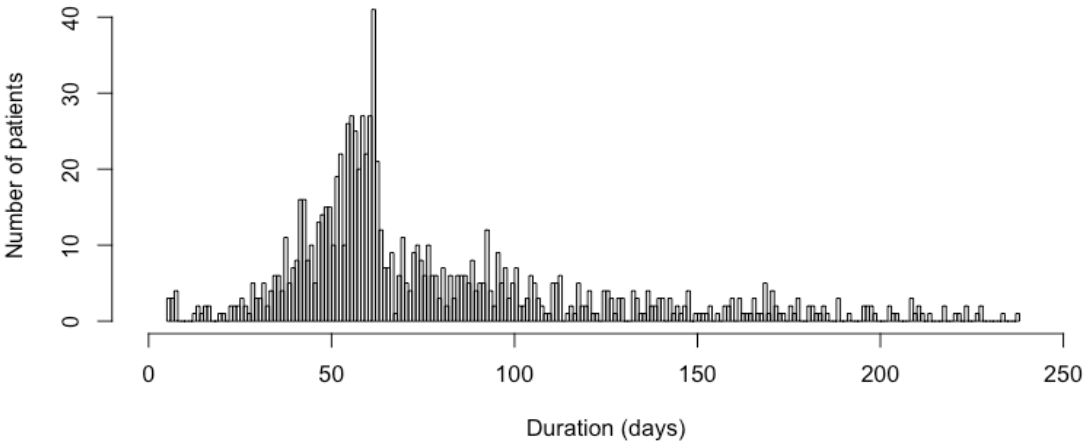


Figure 6.8 Histogram of the duration of endometrial cancer 62-day pathways

The process model, trace variants, dotted chart, and the histogram of the pathway duration of the 62-day pathway of endometrial cancer are the supporting evidence of the complexity of the pathway. The variability is the focus of the process analytics.

The ***process analytics*** step was done by analysing pathway variations based on the first treatment, the age at diagnosis, and the year of diagnosis. The groupings were shown in Table 6.5. For each grouping, a comparison was done to the pathway duration and the percentage of activity presence in the traces.

Based on the ***first cancer treatment***, there are three groups: *Chemotherapy*, *Radiotherapy*, and *Surgery*. Variability of the treatment duration in those three groups was compared. The shortest median duration is *Chemotherapy* (min 15 days, median 60 days, max 228 days). This is followed by *Surgery* (min 5 days, med 62 days, and max 233 days). The *Radiotherapy* group (min 27 days, med 97 days, and max 238 days) had the longest median duration among the three. *Surgery* is the most common treatment with a median duration similar to the target duration (62 days).

The next comparison was done using the Process Comparator plugin in ProM. This plugin compares the percentage of trace frequencies having activities over all patients within a group. The alpha significant level is 5%. A summary of the differences between two groups is presented in Table 6.6. Based on the ***age at diagnosis***, the patients were grouped into their 20s, 30s, 40s, 50s, 60, 70s, 80s, and 90s. The comparison was done by comparing the duration of treatment within the group of patients. The results show no specific trend on the treatment duration based on the age at diagnosis. The four most frequent groups have a very similar distribution of treatment duration, which are the 50s, 60s, 70s, and 80s groups.

Table 6.6 Process comparison based on the first treatment

Group 1	Group 2	Difference	Differences (Group 1 : Group 2)
Surgery	Chemotherapy	14.08%	*Outpatient* (40% : 71%) *Consultation* (55% : 79%) *Discharge* (40% : 75%)
Surgery	Radiotherapy	11.11%	*Admission* (93% : 71%) *Outpatient* (40% : 59%) *Discharge* (40% : 63%)
Chemotherapy	Radiotherapy	4.6%	-

The next comparison was using the Process Comparator plugin. A summary of the pair-wise differences is presented in Table 6.7. This shows that the differences between two groups range from 0% to 10% (average 3.6%). The group with the most significant difference is the 90s group (average 7.3%). Based on age at diagnosis, patients can be grouped into their 20s, 30s–40s, 50s–80s, and 90s.

Table 6.7 Pair-wise differences (%) based on the age at diagnosis. Percentage is color-coded from red as the highest to green as the lowest difference.

	20s	30s	40s	50s	60s	70s	80s	90s
20s								
30s	2.63							
40s	3.17	0						
50s	3.39	3.28	1.64					
60s	3.33	3.33	3.33	0				
70s	3.33	1.67	3.23	5.17	0			
80s	2.99	2.94	1.56	3.28	1.64	0		
90s	2.7	4.76	6.67	8.47	10	10	8.82	

Based on the ***year of diagnosis***, the patients were grouped into 2003 to 2017. The comparison was done based on the treatment duration. The result shows the variability of the treatment duration based on the year of diagnosis. The trend of the median duration as shown in Figure 6.9 increased in 2003 (84 days) to 2005 (97 days). It decreased slightly in 2006 (66 days) to 2012 (57 days). It then increased in 2013 (72 days) followed by a decrease in 2014 (61 days) to 2017 (60 days).

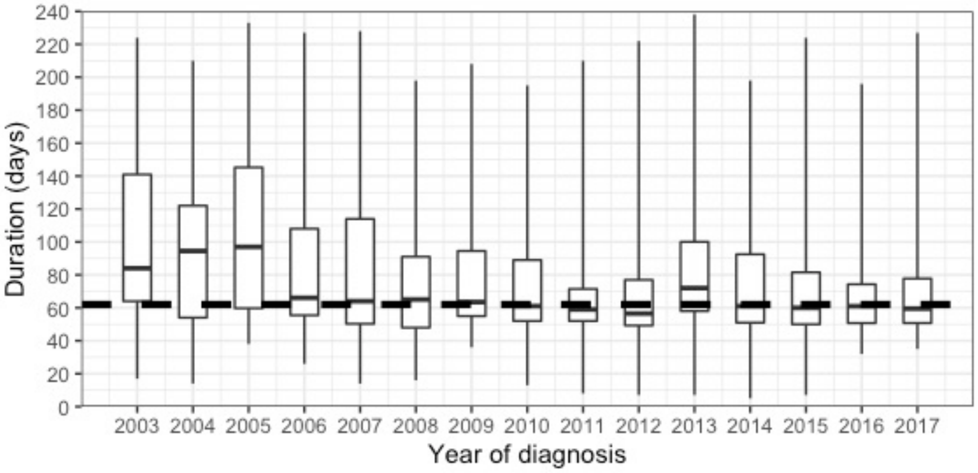


Figure 6.9 The treatment duration based on the year of diagnosis. A box shows the first quartile, median, and the third quartile. The lines show the variability from the minimum to the maximum duration. The dashed line shows target duration (62 days).

6.4.4 Stage 4: Evaluation

In this experiment, the evaluation was done using both statistical and clinical evaluation. The results of the statistical evaluation were presented along with each result in Stage 3. The other detailed results of this experiment are presented in the paper prepared for a journal. The results of the clinical evaluation in each stage of this experiment are summarised in the following paragraphs.

In ***Stage 1***, planning and justification was done as discussed with clinical experts. They suggested grouping the patients to explore the variations in cancer treatment. The resulted patient groups were based on the first treatment, the age at diagnosis, and the year of diagnosis.

In ***Stage 2***, discussions with clinical experts were conducted to validate the 12 tables selected in this study. The clinical experts also suggested a limitation on the GP referrals as the start event to include only referrals from the GP to one of four oncology specialisms. Those four specialisms are gynaecology, gynaecology oncology, medical oncology, and clinical oncology. Another input from the clinical experts was to separate diagnostic and therapeutic surgeries. The last input from the clinical experts in this stage was to limit the pathway duration up to 240 days. This was based on the real practice that, when they need to diagnose a patient, they consider events up to eight months before the diagnosis only. A longer time period is too long for the events to be related. In the transformation step**,** because the database contained timestamps in the granularity of days, some events seemed to happen in parallel when they were recorded on the same day. This was handled by setting an order based on the logical sequence provided by clinical experts.

In ***Stage 3***, clinical expert feedback was that the visualisations resulting from this experiment are interesting and useful for further analysis of the pathways. Some limitations raised for each visualisation type (process model, dotted chart, boxplot diagram) were focused on that fact that there is not one visualisation to fit all requirements within the analysis. This was reflected as an opportunity to improve the visualisation approaches in process mining. A potential future work is to list a set of process mining visualisation methods along with their features and a list of frequently-posed questions in pathway analysis. Those two lists could then be mapped to build a guidance on the recommended visualisation methods for different questions of pathway analysis.

This experiment provides evidence that the PPM Cancer dataset contains sufficient data for analysing cancer pathways from GP referral to the first treatment. The question-driven methodology has been applied and is suitable for an exploratory study to analyse variants in a treatment pathway.

6.5 Experiment 8: Cancer treatment change analysis

This experiment was done to detect change points in the monthly records of events related to cancer treatments. The experiment was to show a change pattern in the system usage and aims to find change points where the system has been potentially changed. It is important to note that this experiment was conducted only for the larger cohort of all cancer patients. This was done to get a larger dataset that can be flexibly partitioned into subsets based on the monthly records. It is not possible to do this experiment in a cohort of patients having a specific type of cancer without breaching the rule of small number limitations. For example, as presented in Table 6.5, the number of endometrial cancer patients over the years ranged from 29 to 105 patients. If that number of patients were then partitioned into months, there would only be 2 to 9 patients per month. Those numbers are too small and patients might be identified.

6.5.1 Stage 1: Planning and justification

The data for this experiment included all events of cancer patients in the PPM Cancer dataset. The analysis was done to examine the number of monthly records of each activity of interest and to detect change points based on the pattern of the monthly records. The monthly duration was chosen to get a smaller size of sub-logs without breaching the information governance requirement to work on small size samples. The research question was "*Is it possible to detect change points based on the monthly records over time?*". This experiment included a team of computer scientists, clinical experts, and statisticians.

6.5.2 Stage 2: Extraction, transformation, and loading

The ***extraction*** was done by a query to get the number of monthly records of each activity related to cancer treatment in the PPM Cancer dataset from 2003 to 2018. The events were recorded in 12 tables, as identified in Section 6.1.3.

The ***transformation*** was done to transform the number of monthly records of each activity from 2003 to 2018 into a time-series object in R. This section presents the step-by-step analysis of the *Diagnosis* events and summarises the findings of the other events. Diagnosis is the baseline in this experiment because it was used in the selection criteria. Patients were included in this experiment if they have been diagnosed with cancer. The complete result of each event is shown in Appendix E.3.

6.5.3 Stage 3: Mining and analysis

The ***process mining*** step in this stage was done by creating a process model of cancer treatment in the PPM Cancer dataset. The process model was very complicated and, as such, is usually called a spaghetti model.

The ***process analytics*** was done to analyse process change over time. The time-series object was then decomposed using the signal decomposition method for additive time series. This method resulted in four plots mapping:

(1) *the observed signal* showing the monthly records of an activity in the event log,
(2) *the trend signal* showing the increasing/ decreasing pattern of the observed plot,
(3) *the seasonal signal* showing the monthly average of the observed signal, and
(4) *the random/ residual signal* showing the residual signal from the observed signal minus the trend and the seasonal pattern.

For example, the result of the *Diagnosis* event is presented in Figure 6.10.

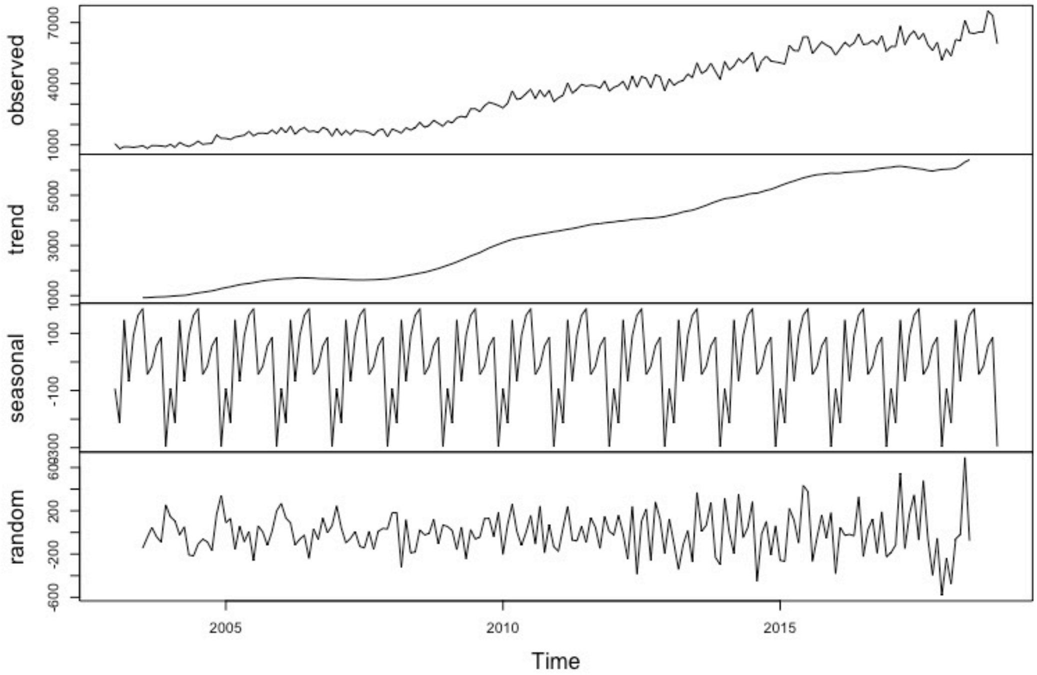


Figure 6.10 Decomposition of monthly records of diagnosis. It consists of four plots mapping the observed, trend, seasonal, and random/ residual plots over time.

The first plot in Figure 6.10 represents the *observed* signal as the monthly records of diagnosis with minimum 807, median 3,641, mean 3,512, and maximum 7,568. The *trend* is presented in the second plot and was based on 12-moving average. The trend can also be explained by fitting a linear model on the observed signal. The fitted linear model for monthly records of diagnosis is presented in Figure 6.11.

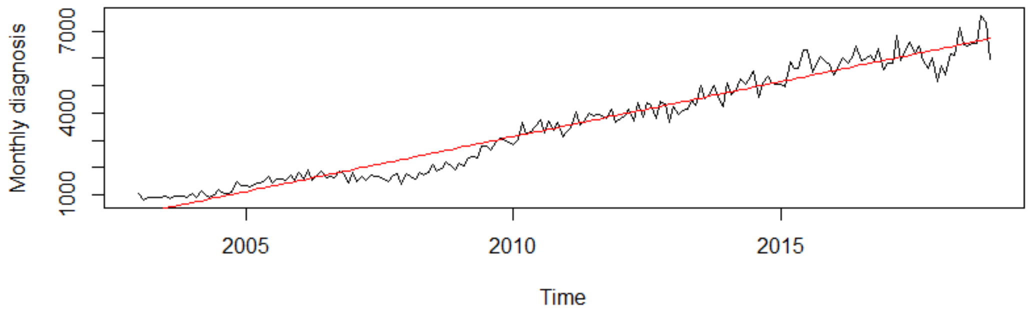


Figure 6.11 Fitted linear model of monthly records of diagnosis. The monthly records increased over the years with a coefficient of 33.6.

The *seasonal* pattern as shown in the third plot of Figure 6.9 is the monthly average of the observed signal. It presents the variability of the number of diagnosis records in a year. The minimum is in December (–296) and the maximum is in July (186).

The *residual/ random* signal is presented in the fourth plot. The residual signal was analysed using a Statistical Process Control (SPC) approach. The idea is to plot the variability of the monthly diagnosis records after subtracting the trend and seasonal patterns. The change points were detected as the residual signal varied outside the control lines. Figure 6.12 shows the SPC chart of the monthly diagnosis records.

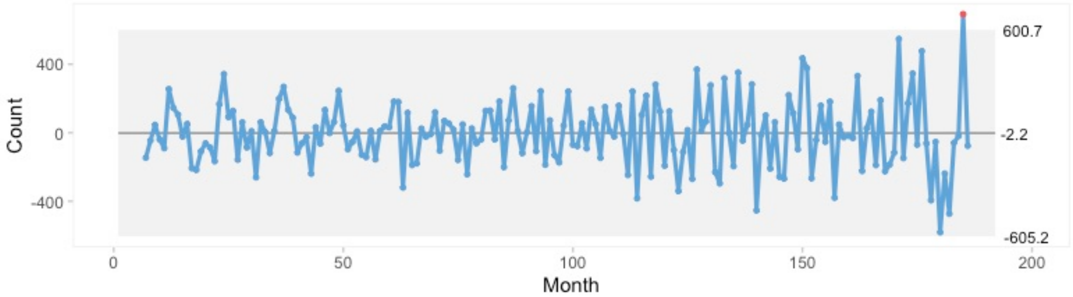


Figure 6.12 The SPC of residual signal of monthly diagnosis records. The blue signal shows the variance over means of the monthly records. One detected change point is in May 2018 where the signal is over the upper control line.

The steps of analysing monthly records with the signal decomposition and SPC methods were performed on each of the 12 activities. Details of the steps in those 12 activities are presented in Appendix E.3. The results were summarised in the statistical evaluation and discussed with clinical and technical experts.

6.5.4 Stage 4: Evaluation

The evaluation was done in two steps: statistical evaluation and clinical evaluation. The statistical evaluation was done by summarising the results of the steps in Stage 3 for each activity. The clinical evaluation summarises the feedback from the clinical experts about the findings.

The ***statistical evaluation*** started with a summary of the fitted linear model and is presented in Table 6.8. The p-value of the fitted linear model indicates the relationship between the number of monthly records and time. A small p-value means it is unlikely that the relationship is due to chance. The coefficients are colour coded to easily spot the activities having the highest coefficient to the lowest. The activity with the highest coefficient is *Outpatient*, and the lowest is *Radiotherapy*. All activities have a small p-value, except *Radiotherapy*. The p-value of *Radiotherapy* is more than 0.05, which means that there is not enough evidence to declare a relationship between the number of *Radiotherapy* records and time.

Table 6.8 Summary of the fitted linear model of the trend pattern

activity	coef	adjusted R^2	p-value
Discharge	113.6616	0.8146	< 2.2e-16
Consultation	72.78571	0.7631	< 2.2e-16
Chemotherapy	12.6943	0.9111	< 2.2e-16
Diagnosis	33.59295	0.9566	< 2.2e-16
Investigation	376.0788	0.9156	< 2.2e-16
MDT Review	39.9816	0.9678	< 2.2e-16
Outpatient	791.9042	0.7555	< 2.2e-16
Pathology	24.8572	0.6626	< 2.2e-16
Radiotherapy	0.1599	0.002322	0.2317
Referral	222.2942	0.8401	< 2.2e-16
Surgery	41.24259	0.8681	< 2.2e-16

A summary of the ***seasonal*** pattern is presented in Table 6.9. Based on this table, activities are mostly at their minimum average in December, except for *Consultation* and *Chemotherapy*. *Consultation* was at the minimum average in August while *Chemotherapy* was at the minimum average in February.

Table 6.9 Summary of the seasonal pattern

Activity	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
Admissions	525	-1052	1131	-728	287	194	621	-266	63	429	147	-1350
Discharge	237	-1048	1130	-631	304	181	690	-215	-54	491	72	-1157
Consultation	4	134	354	-100	-28	415	98	-650	275	-425	91	-167
Chemotherapy	19	-161	48	-41	-4	4	54	40	26	45	17	-47
Diagnosis	-93	-214	147	-68	92	163	186	-44	-15	56	86	-296
Investigation	2468	-1266	2452	-688	463	179	1405	-1490	-140	1131	120	-4634
MDT Review	101	-177	50	-146	-67	162	224	-128	45	135	107	-306
Outpatient	3200	-8416	1939	-2370	770	5126	4956	-5704	3862	5730	4474	-13567
Pathology	10	-253	222	-172	8	128	51	-232	147	212	320	-440
Radiotherapy	4	-48	2	-42	-30	8	12	-22	5	101	76	-66
Referral	802	-1348	2154	-377	601	1438	1262	-1739	204	1155	-88	-4064
Surgery	26	-318	270	-279	-56	215	170	-241	177	284	346	-594

Key: Red = Maximum, Yellow = Minimum

Statistics of the seasonal pattern is presented in Table 6.10. The variability of the seasonal pattern can be described based on the interquartile range (IQR) value. An IQR represents the middle 50% of values ordered from the lowest to the highest. The activity with the highest IQR is *Outpatient* with 7,798 and the lowest IQR is *Radiotherapy* with 38.1.

Table 6.10 Statistics of the seasonal pattern

Activity	Min.	Q1	Med.	Q3	Max.	IQR
Admissions	-1349.8	-381.6	170.4	452.7	1130.8	834.3
Discharge	-1157.3	-319.4	126.4	351.0	1130.0	670.4
Consultation	-650.4	-116.8	47.5	169.0	415.1	285.8
Chemotherapy	-161.1	-13.1	18.0	41.2	53.8	54.3
Diagnosis	-295.9	-74.4	20.5	106.0	186.3	180.4
Investigation	-4633.6	-832.5	149.6	1199.5	2467.7	2032
MDT Review	-305.5	-132.5	47.2	113.8	223.9	246.3
Outpatient	-13567.0	-3203.6	2569.5	4594.4	5729.6	**7798**
Pathology	-440.3	-187.2	30.6	163.1	319.9	350.3
Radiotherapy	-65.7	-30.1	3.4	8.0	101.0	**38.1**
Referral	-4064.1	-376.6	600.5	1235.3	2153.7	1611.9
Surgery	-593.8	-241.4	98.4	214.8	345.8	456.2

A summary of the detected change points based on the *random/residual* pattern is presented in Figure 6.13. January 2004 was detected as a change point based on five activities (*Admission, Discharge, Investigation, Outpatient*, and *Referral*). It was followed by February 2004 based on four activities (*Admission, Discharge, Investigation,* and *Outpatient*), October–November 2003 based on three activities (*Admission, Discharge*, and *Investigation*) and October 2014 based on three activities (*Admission, Discharge*, and *Radiotherapy*).

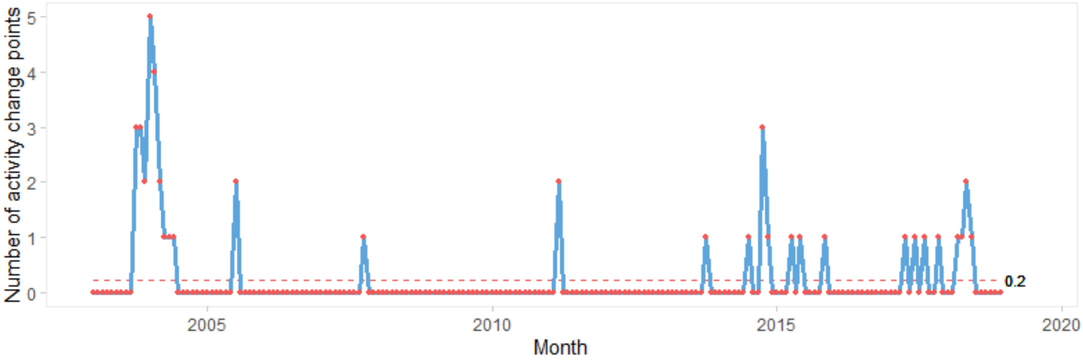


Figure 6.13 Number of activity change points by month. It shows the number of activities where change points were detected, over time.

The ***clinical evaluation*** was done through a discussion of the findings of this experiment. Some comments from the clinical experts were related to each finding and were mostly on the change points detected from the residual analysis using SPC. The details of the analysis of each event are presented in Appendix E.3. Some important comments are summarised as follow.

The *Admission* and *Discharge* are two related activities which record the hospital admission and discharge of a patient. The pattern of detected change points based on the admission and discharge are very similar. Both suggesting three change points in 2004, 2011, and 2014. Discussion with the clinical experts revealed that admissions were recorded as a copy from another system in the early years of the PPM system (2003–2004) and those had been completely migrated in later years.

The *Consultation* shows change points detected in March–June 2018. This was discussed with the clinical experts and was apparently related to the fact that *Consultation* was moved into another system recently and was in the process of migration during 2018.

The *Chemotherapy* shows a change point detected in July 2014. Based on discussion with the clinical experts, this is likely due to the change when clinicians in the haematology department started to do chemotherapy, whereas previously this could only be done by clinicians in the gynaecology and gynaecology oncology departments.

Across activities, there are two periods where changes had potentially happened: October 2003-January 2004 and October 2014. October 2003-January 2004 was the early period of PPM implementation. Further discussions revealed that the change in October 2014 was due to the launching of the Leeds Care Record initiative.

6.6 Summary

This chapter has presented four experiments using the PPM Cancer dataset. The first one analysed the pathway from referral to the diagnosis of endometrial cancer. This has been published in a conference journal in 2019 [177]. The second one analysed the 62-day pathway of endometrial cancer, which is the pathway from referral to the first treatment of cancer. This is prepared for a journal paper [178]. The third one presented the analysis of a GP Tab as a feature in the PPM EHR system. The GP Tab was treated as a known change in the system and was analysed to explore how this feature had been used during chemotherapy cycles. The fourth one explored the records in the PPM EHR system to detect change points down to the month. The case study of PPM Cancer dataset can also be seen from a different perspective, as described in the following paragraphs.

From the health service perspective, process models discovered using process mining approaches are useful to visualise the pathways and further analysis. For example, experiment 6 and 7 in this case study can express process models in several visualisations, such as directly-follows graph in Figure 6.6. Future analysis was done following multi-level approach to analyse pathways in process-model, trace, and activity levels.

From the process mining perspective, this case study has been shown that the PPM Cancer data can be used for process mining in healthcare. As a representative of real-life healthcare data, the PPM Cancer dataset has been shown to provide sufficient data needed in process mining. The limitation of using process mining has been presented in Section 6.1.6. Due to the high number of data and variations of those data, a straight-forward approach to use process mining has not resulted in useful insight about the process. This has been addressed by proposing a combined approach of signal decomposition and SPC chart.

From the information system perspective, this case study has been shown two experiments to analyse process change over time, which are experiment 5 and experiment 8. Experiment 5 shows how GP tab change has been used as an example of a known change in the system. The analysis has been done by creating an event log by combining user access log in the PPM Splunk and patient log in the PPM Query database. Experiment 8 proposed a new approach to analyse process change over time based on signal decomposition and SPC chart. The detected change points were then being discussed with experts to find the potential reasons behind those changes.

The next chapter summarises discussions based on Chapters 1–6 of this thesis. It serves as a critical exploration of the literature review and method development explored in this study and as a reflection on the experiments conducted in the three case studies described in Chapters 4–6.

Chapter 7
Discussion

Chapters 4 to 6 presented the analysis of the case studies based on the methodology described in Chapter 3. In this chapter, a discussion on the findings and lessons learned in this study is presented. This includes a reflection on the method, the challenges with healthcare process mining, the process change analysis, and the effect of system change on the healthcare process. This chapter is concluded by a discussion on the contributions of this thesis.

7.1 Reflection on the method

The standard method presented in Chapter 3 has been applied to the three case studies. The method evolved and was refined, as presented in Figure 7.1.

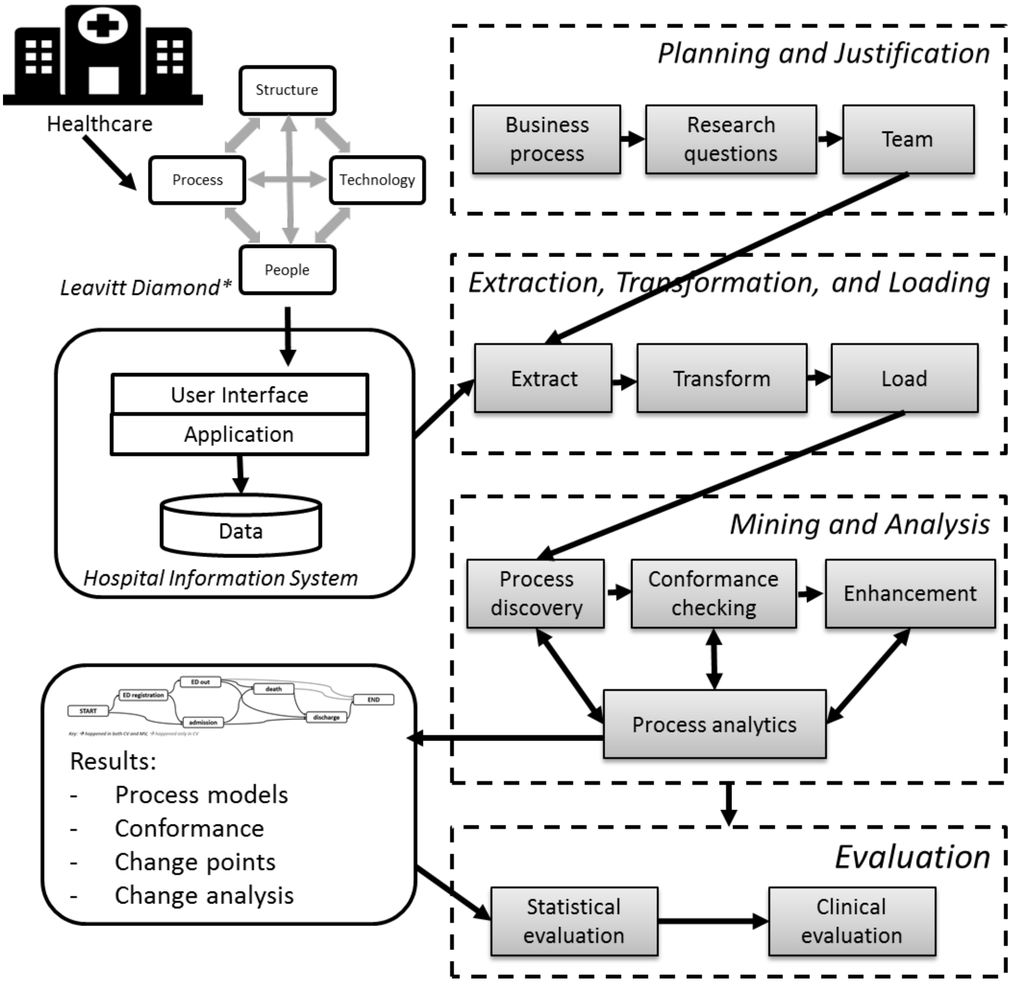


Figure 7.1 Reflection on the method. The final method (right) has been put into the context of the key inputs and outputs in the healthcare environment (left).

As presented in Figure 7.1, the healthcare process in reality is complicated by the multifaceted nature of the organisational structure, process, people, and technology. These are also known as four forces in Leavitt's diamond, as discussed in Section 2.2.4.2. All of these facets are captured by the Electronic Healthcare System (EHR) through its User Interface (UI). They are then processed by the application and recorded in the database of the EHR. The data in the database was extracted, transformed, loaded (ETL), mined and analysed in this study, and later reviewed based on both a statistical and clinical evaluation.

The benefit of process-oriented data analysis is that it uses the routinely collected data within the EHR. In terms of research, it means that there is no additional process or effort required for data collection. Another advantage is that the analysis can be done in a relatively short time over a relatively long duration of the data. By using the routinely collected data within the EHR, the process model discovered reflects the real execution of the treatment process.

7.2 Answering research questions

The six Research Questions (RQs) as presented in Section 1.3 have been broken down and answered through experiments using three datasets. The summary is as follow.

(***RQ-1***) It is possible to analyse changes in care processes in EHR using process mining. This has been answered through experiment 2 as presented in Section 4.3, experiment 3 in Section 5.2, and all four experiments using the PPM Cancer dataset in Sections 6.2 – 6.5. This research questions covered RQ-2 and RQ-3.

(***RQ-2***) It is possible to analyse process changes from a given UI change. This has been answered in experiment 2 in Section 4.3 and experiment 5 in Section 6.2. The MIMIC-III dataset provides a case where the EHR system was changed from a version to another. Process changes have been analysed by comparing the trace of the process before and after the system change. The analysis of a UI change in the PPM Cancer dataset is represented by an introduction of GP tab as a new feature in the PPM EHR system. Process mining approach has been used to analyse the usage of the new feature within the care pathway of chemotherapy cycles.

(***RQ-3***) It is possible to detect a time point when a care process changed and this has been shown through experiment 6 – 8 in Sections 6.3 – 6.5. Experiment 6 analysed process changes in the pathways from GP referral to the diagnosis of endometrial cancer. The multi-level approach applied in this experiment found several change points in the pathway. Experiment 7 explored the variability of the pathway from GP referral to the first treatment of endometrial cancer and the 62-day target. The variability of the pathway was explored based on the type of the first treatment, the age at diagnosis, and the year of diagnosis. Experiment 8 analysed the process change in the dataset of all cancer patients in the PPM system. Some change points were detected and reflected different causes of process change, including the early stage of EHR system implementation, the introduction of a new system to record an event in the EHR system, and the change in the structural organisation to the upgraded capacity of the care service in the hospital.

(***RQ-4***) A process change is characterised by many parameters. This has been explored in experiment 4 in Section 5.3 and in experiment 6 through the multi-level approach in Section 6.3. Those parameters include replay fitness, precision and generalisation in the process-model level; trace duration and variant proportion in the trace-level; and frequency and percentage of activities in the activity level.

(***RQ-5***) The best representation of care pathways depends on different perspectives for analysing care pathways and the properties of different options have been explored in Section 3.2.3.1. The options are to visualise the care pathway as a process model, a trace variant diagram, a dotted chart, or a process comparison diagram, among others. All experiments in this research explored those options and presented the best ones to support the visualisation of the experiment results.

(***RQ-6***) The extraction of the dataset of patient pathways from the EHR system can be done by identifying events related to the patient pathways, extracting those events and transform them into an event log with minimum components of *case id, activity name,* and *timestamp*. Those steps represented stage 1 (planning and justification) and stage 2 (extraction, transformation, and loading) in the study method.

Some challenges in working with healthcare data have been reviewed in Section 2.2.4.2 and are described in more detail in Section 7.3. The analysis in this research was heavily dependent on the data quality, which is dependent on many aspects of the collection, design, structure, management, and policy of the data. The quality of the analysis results was also dependent on the ETL, tools, method, and clinical experts. Another limitation is that the discovered process models can only represent what has been recorded in the data.

7.3 Challenges on working with healthcare data

During this study, there were some challenges identified in performing process mining using healthcare data. The main challenges were related to data access and ethics approval, data quality, data understanding, and data visualisation. These will be discussed in the following sections.

7.3.1 Data access and ethics approval

Healthcare data are personal and sensitive. The procedures to gain access to the data varied based on the institutions managing the data. In this study, three datasets were used, and three different approaches were undertaken to gain access to the data.

The MIMIC-III database is accessible after completing the National Institutes of Health (NIH) Protecting Human Research Participants training course. Upon completion, the certificate (as presented in Appendix B.1) had to be uploaded along with the MIMIC-III access request through a PhysioNetWorks login account. Another requirement was the need to sign a data use agreement. When access is granted, the MIMIC-III database can be downloaded as comma-separated (.csv) files. Those files can be imported to any major database frameworks such as MySQL, PostgreSQL, and Oracle. In this study, PostgreSQL was used.

The PPM Chemotherapy dataset was used in two previous studies as described in Section 3.4.2. The front page of the IRAS application is presented in Appendix B.2. Access to the PPM Chemotherapy dataset was gained by joining the research team working on this dataset. This process was accompanied by a series of discussions with the PPM developer team to support connectivity setting and data understanding. Some challenges were solved through those discussions, for example, we had to decide whether data would be accessed from a University-networked PC or a hospital-based PC. We also submitted a request to create a Virtual Research Environment (VRE) in the Leeds Institute of Data Analytics (LIDA), but that was not successful. Once the decision was made, i.e. to enable access using a remote desktop connection from a University-networked PC, the process was completed within a week.

The PPM Cancer dataset is an extract of the PPM live database, as described in Section 3.4.3. In this study, access to the database was through a hospital-networked PC in a LIDA secure room. The access arrangement to the PPM Cancer dataset was made through two approaches which were the LTHT Honorary Contract and the

Integrated Research Application System (IRAS) application. The LTHT Honorary Contract was first submitted in July 2016 and access granted in December 2016 (a five-month processing time). The IRAS application took longer. It was first created in April 2016, progressed and finally submitted on November 2017, and the approval outcome was granted in April 2018 (a two year processing time). Access to the PPM Cancer dataset began in January 2017, and access to the complete data was given in March 2018. The processing time of the ethics approval delayed the progress of the experiments with the PPM Cancer dataset.

7.3.2 Data quality

Data quality is a critical challenge in data analytics, including healthcare process mining. The EHR data were collected for clinical purposes to support clinicians providing treatment for their patients. When those EHR data are used for research purposes, the suitability of the data may be compromised. This condition may result in an extensive amount of data cleaning and filtering needed for the research.

In this study, the data quality of each dataset was assessed following the Weiskopf & Weng framework [133]. This framework specifies five data quality dimensions: completeness, correctness, concordance, plausibility, and currency. Those five dimensions were assessed using seven methods: element presence, data element agreement, data source agreement, distribution comparison, validity check, gold standard, and log review. The main purpose of the data quality assessment in this research was to assess its suitability for process mining. Data quality of the datasets for process mining has been discussed in Sections 4.1.3 (MIMIC-III), 5.1.3 (PPM Chemotherapy), and 6.1.4 (PPM Cancer).

The main data quality issues for process mining in healthcare can be identified as incompleteness, incorrectness, impreciseness, or irrelevancy [104]. These can happen in different levels of the event log: case, event, attribute, activity names, timestamp, or resource. The general approach to handle data quality issues was to identify the issues and exclude the unsolved issues. The advantage is that this approach resulted in high-quality data for further analysis. The disadvantage is that the resulting data was relatively small and might not representative of the complete dataset. This is the only possible approach within the MIMIC-III data analysis because access to the hospital is not possible. In the PPM Chemotherapy and PPM Cancer datasets, this approach can be minimised through frequent discussions with the clinical experts.

Another lesson learned from analysing the three datasets in this study was that the data quality was continually changing over time, due to many changes happening within the system. In the MIMIC-III dataset, the system change from CareVue (CV) to MetaVision (MV) was the system change of interest. Process comparison in the CV and MV systems has been discussed in Section 4.3. In the PPM Chemotherapy and PPM Cancer datasets, the data quality changes were found when comparing the process execution over time. Conformance-based change detection in PPM Chemotherapy dataset has been presented in Section 5.2.3. In the PPM Cancer dataset, process change has been analysed based on a known change in Section 6.2 and without a known change in Section 6.3 – 6.5.

7.3.3 Data understanding

Data understanding is another challenge in analysing healthcare data. This challenge is related to many factors, such as coding standards, semantic meaning, and also data quality. Understanding the nature of the EHR system, coding standards, process guidelines, the nature of cancer as a complicated disease, and UK standard for cancer waiting times have been effected by reviewing the related literature, as presented in Section 2.1. The nature of healthcare data as 'big data' [181] increased the challenge to understand the data completely. Big data can be described in the five *V*s. The *Volume* of big healthcare data refers to the vast amount of data recorded in the system. The *Velocity* refers to the speed at which new data is generated and the speed at which data moves around. The *Variety* refers to the different types of data in the HIS, which includes among others administrative data, clinical data, radiology, and imaging. The *Veracity* refers to the messiness or trustworthiness of the data. The *Value* refers to the usefulness of the data in the specific domain.

Based on the analysis of the three datasets in this study, this challenge was related to the different conditions within the three different datasets. The understanding of the MIMIC-III dataset was done completely based on the available documentation, both from the website of the MIMIC-III dataset and published papers analysing the MIMIC-III dataset for many other purposes. This understanding was limited because there was no direct access to the hospital that provides the data. Another limitation was due to the anonymisation approach undertaken by the data curator of the MIMIC-III dataset, where times were shifted to the future dates. Some analysis was not possible, such as the analysis of busy days, the impact of a bottleneck, etc. This has been presented in Section 4.1.6.

The understanding of the PPM Chemotherapy and PPM Cancer datasets was done based on the available documentation, data exploration, and through discussion with the clinical experts and the development team. The documentation of the data structure of the PPM Chemotherapy and PPM Cancer datasets is limited. Direct exploration of the data is therefore required and this, in turn, requires a significant amount of time. The documentation of the system change in the PPM Cancer dataset was provided through access to PPM JIRA. In the PPM JIRA, all records of system changes can be found, including the change request, steps undertaken in the execution of the change, and the release notes. Further discussion with the clinical experts and the development team were found to be useful to complete the understanding.

7.3.4 Data and process visualisation

This study involved both computer science and clinical insights. Knowledge transfers between those two disciplines were completed through several discussions. Thus, a good visualisation approach to support the discussion was required. Basic data visualisations to describe different perspectives of the dataset have been used in this study, such as bar charts, line charts, histograms, box-and-whisker plots.

During the analysis, some visualisation approaches were used to facilitate discussions with the clinical experts. The results of the process mining have been presented in several complementary ways. The challenge was to decide which one of the visualisations can be used to present the results in the best way. As discussed in Section 3.2.3.1, the main visualisations are process models, trace variants, dotted charts, and comparison diagrams.

A process model is the main output of process discovery in process mining. A process model can be presented in several diagrams, such as a transition system, a UML activity diagram, a BPMN model, or a Petri Net as described in Section 2.2.2. In general, a process model is a diagram showing the sequence and the flow from one activity to another. Additional information can be presented in a process model such as the number of traces and the median or average time between events. Colouring and thickness can also be used to show the frequency or importance of the event and the path from one event to another.

A trace variant diagram shows the sequence of events happening to traces, grouped by the trace variants. An example of a trace variant diagram is shown in Figure 4.5 in Section 4.2.4. It is normally sorted in descending order of the frequency of the traces in the trace variants. This visualisation is useful to see the variability of the pathways. Limitations of a trace variant diagram are that it is not representing the duration of activities in the process and that one variant is represented as one sequence without representing the volume/ proportion of the variant relative to the population.

A dotted chart shows the spread of activities over time. An example of a dotted chart is shown in Figure 5.5 in Section 5.2.3. The most common setting is to have a dotted chart with time on the x-axis and patients on the y-axis. One line in the dotted chart shows the pathway of a patient over time. A dotted chart is useful to gain insight on the pathways from the time perspective. A limitation of a dotted chart is the difficulty to represent the general pathway of a group of patients. This limitation is because one dot in the dotted chart represents one event happening to one patient, giving a fine-grained level of pathway.

A comparison diagram is a state transition diagram showing activities as boxes and the paths between activities as directed arcs from one box to another. In these terms, a comparison diagram can also be seen as a process model. The difference is that a comparison diagram represents a process model from the comparison of two event logs. It is then annotated and colour-coded to visualise the differences in those two event logs. The advantage of using a comparison diagram is that it combines process models from two event logs being compared as one diagram. The limitation is that the combined process model increases complexity.

7.3.5 Process change over time

The main challenge discussed in this study is the healthcare process changes over time. Any analysis done in healthcare should consider those changes and how they affect the process of interest. This was also the case in the healthcare process mining projects. It is important to collect and analyse a large amount of data to get a better result for process mining. The trade-off is that a large amount of data are mostly collected during a long period of time. The EHR system collecting those data might have been changed. If the data is assumed to be static, the process model and the analysis of it would be invalid. The conformance of the discovered process model needs to be measured over time. In the first and second experiments of the PPM

Cancer dataset case study, it was evidenced that the process has been changed over time. This change was analysed with the multilevel approach proposed in this study.

Data in an EHR system is an output of a complex combination of organisational structure, task, people, and technology [135]. A change in an EHR system could be a change in organisational structure, task, people technology, or a combination of those aspects. The specific cause of a change was not discussed in this study. A change in the EHR system is an interplay between those aspects. For example, when there is a change in the organisational structure, the system will need to change the task, people, and technology. The detected changes were discussed with the clinical experts to reveal the possible reasons behind those changes, but there was no specific method to differentiate change in the organisational structure or the task, people, or technology.

The process change analysis in the MIMIC-III was effected based on a system change documented in the MIMIC-III data descriptor. This change could be seen as a change in technology. The findings showed that the organisational structure, task, and people were also changed. This has been discussed further in Section 4.3.5.

The analysis in the PPM Chemotherapy dataset was done by comparing process models over years, based on the value of trace fitness, precision, and generalisation. This analysis found differences in the process models, that could be used to represent processes or task. This has been discussed further in Section 5.2.4.

The analysis in the PPM Cancer dataset was achieved in two ways: with or without a known change prior to the analysis. With a known change, analysis of the effect of GP tab change showed an example of a technology change. The finding showed the evidence of organisational-, task-, and people- changes. Without a known change, the findings showed evidence of changes in all four aspects. As discussed in Section 6.5.4, it revealed that the technology has changed over time, from a software focused on supporting cancer data collection to an EHR system in the whole hospital trust, from a software coded in Basic to a web-based application. The EHR system is also evolving, with some organisational-, task- and people- changes found where the activities were started to be recorded in the system at different years, additional task was given to a department, and organisational change to join a larger initiative combining the EHR system with those of other institutions.

7.4 Process change analysis

In this thesis, one of the main tasks was process change analysis. This analysis was done by partitioning an event log to create subsets over time and performing a process comparison between two consecutive time windows. The partitioning approach was applied, as illustrated in Figure 3.8. The general method of using process mining to analyse process changes has been built upon the available approaches in the literature including concept drift analysis, as described in Section 2.2.5. The problem dimensions of process change analysis are illustrated in Figure 7.2 below.

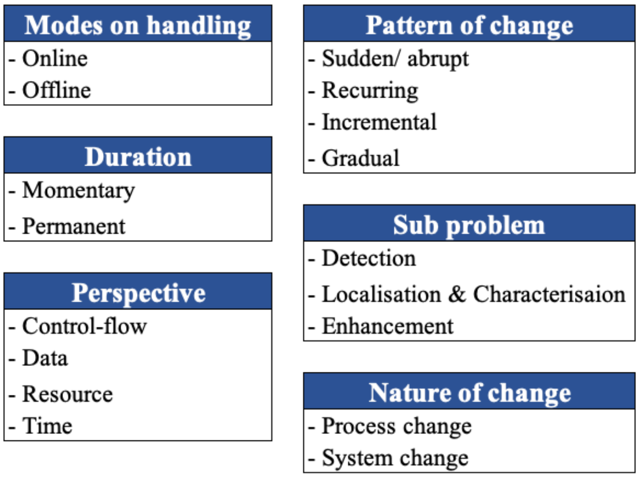


Figure 7.2 Problem dimensions of process change analysis. Each dimension needs to be analysed in the initial stage of process change analysis.

It is important to understand as many dimensions as possible in the initial stage of process change analysis, to plan and justify the method correctly. Each of the problem dimensions in Figure 7.2 is described in the following paragraphs.

In term of the *modes of handling*, a process change analysis can be done in an online or offline mode. This study was in the offline mode, which means that the process change analysis was not performed on the real-life/real-time data but was performed on the extracted data. An unusual condition on the three datasets used in this study was that access to the datasets was provided to the full database. This condition is advantageous in that it enabled further extractions as required in each experiment.

Based on the *duration* of change, this study analysed both the momentary and permanent changes. In the MIMIC-III dataset, the system was changed permanently from CV to MV in 2008 with a clear separation of hospital admissions. In the PPM Chemotherapy dataset, no specific system change was analysed, but the data partitioning technique was used to analyse process change over time. A conformance-based comparison was undertaken to compare the conformance of two consecutive years, which reflect the momentary change in the process. In the PPM Cancer dataset, the GP Tab implementation was an example of a permanent change. On the other hand, a momentary change was also explored using the data partitioning technique combined with a multilevel approach of consecutive years comparisons.

Based on the *perspectives*, this study focused on the control-flow perspective, but also analysed the data and time perspectives. The control-flow perspective was used to analyse changes in the sequence and flow of events. The data and time perspectives were used to provide supporting details of the control-flow perspective.

The *pattern of change* analysed in this study was a sudden/abrupt change. The sudden/abrupt change is the basic type of change which might also support the finding of other patterns of changes. When a sudden change happened several times with a repeating pattern, a recurring change is potentially identified. When a sudden change occurred several times with a non-repeating pattern, an incremental change or a gradual change is potentially identified.

The *sub-problems* of the process change explored in this study were the detection, localisation, characterisation, and enhancement/unravelling of the process change. When a change is known, change detection is straightforward. For example, in the MIMIC-III dataset (CV to MV) and in the PPM Cancer dataset when the GP Tab introduction was being analysed. When a change was unknown, change detection was done by uniformly partitioning the data over time and comparing the process execution within each time window. Change localisation and characterisation were done by digging deeper into the process based on the identified change. The enhancement was done through discussions with the clinical experts and the development team to reveal the possible reasons for the process change.

The *nature of the change* revealed that the enhanced sub-problem ranged from the real change of process execution to the change within the EHR system. The change in the MIMIC-III dataset was known to be caused by the change of the EHR system. The changes in the PPM EHR system were more complicated because they were a combination of many changes in the system. Process changes were found through process comparison over time. An example of system changes was that some activities were not found in the early years because they were not recorded yet in the system.

7.5 The effect of system change on healthcare processes

A system change in an EHR system could happen any time and for many different reasons. This research explored some reasons behind a system change. In the MIMIC-III dataset case study, a system change happened in 2008 when the hospital replaced the EHR system with a new one. In the PPM Chemotherapy dataset case study, there were no system changes known initially. The analysis of the process revealed that there were differences in the conformance values over time. This finding suggested that there were process changes. The more complicated condition happened in the PPM Cancer dataset case study. The dataset contains clinical data of the patients over 16 years (2003 to 2018). The PPM EHR systems had begun to be used in the hospital in 2003. Some changes had taken place in those 16 years. In 2010, the PPM1 that was focused on cancer treatment records only was expanded to the PPM that covered the whole Trust. In 2014, there was an introduction of a GP Tab that enabled clinicians in the hospital to check on patient records in the GP system during consultations.

Among the three case studies, the PPM Cancer dataset was the best fit for this study. It was because this dataset was a copy of a real-time database, complete with the real advantages and real limitations. The main benefits were that it represented actual healthcare processes in a busy hospital and that it came with an excellent connection to clinical experts and the development team. The limitations were that the dataset was raw and had not been curated. The PPM Cancer dataset case study was also suitable for this study because there were some documentations of system changes over time, which was used by the developer team to record the progress of system changes and communicate within the team. The access to the complete data of patient treatments, the clinical experts, and the development team of the EHR system were three success factors in analysing process change within a healthcare setting.

There were four experiments to analyse the effect of system change in the healthcare process in the PPM Cancer dataset case study. They were experiments 5-8 in Section 6.2 - 6.5. The experiment on the GP Tab change analysis was to test the effect of a UI change on the process. The process mining approach was a suitable approach to reveal the use of the GP Tab during chemotherapy cycles. This experiment was representative of the process change analysis based on a known system change. On the other hand, experiments 6-7 worked on analysing process change over time on the endometrial cancer pathways, guided by the UK cancer waiting time standard, as described in Section 2.1.7. Experiment 8 was done to detect process change without any prior information of a possible system change. The experiment explored the system usage to detect change point(s) that had possibly occurred within the system.

7.6 Contributions of this thesis

The motivation of this research is to contribute to the community, specifically the Process-Oriented Data Science for Health (PODS4H) community [8]. There are four contributions of this thesis: the case studies for healthcare process mining, the multilevel approach for identifying process change, the signal decomposition approach for change analysis, and the time window selection for process analysis. Each of them will be described in the sections below.

7.6.1 Case studies for healthcare process mining

The first contribution of this study was to apply process mining approaches in three different datasets, i.e. the MIMIC-III, the PPM Chemotherapy, and the PPM Cancer datasets. All three datasets did not come from process-aware information systems. A process-aware information system is a software system that manages and executes processes based on process models. Process mining in process-aware information systems is straightforward, which means that analysis can be done based on the process models referred to in the information system development. Because that was not what happened in the three datasets used in this study, additional steps were needed to assess if the quality of the datasets were sufficient for process mining. These additional steps included the data quality assessments focusing on the completeness and validity of the data for process mining. This assessment was done in the data understanding steps, which led to different challenges within the three different datasets.

The MIMIC-III dataset is representative of a typical database in an EHR system. This dataset is publicly accessible and has been curated following the good practice of sharing sensitive healthcare data. Those are why the MIMIC-III dataset is suitable for data analytics projects focused on method development. In this study, the MIMIC-III dataset was used in the first year of study to focus on the exploration of the process mining techniques available in ProM plugins, the DISCO tool, and R libraries. The MIMIC-III dataset was first used in a process mining project in our research group, including in this study. There is still a wide range of possibilities to use the MIMIC-III dataset for many other studies in process mining.

The PPM Chemotherapy dataset was initially extracted for a study focused on analysing the effect of adverse events in cancer treatments, especially chemotherapy. The previous study was done using specifically developed software and a Markov model to produce a schematic diagram of patient pathways during chemotherapy. This current study improved the method used in the previous study by applying process mining as a structured approach to analyse patient pathways during chemotherapy as a cancer treatment. This dataset is rich with clinical details and is potentially useful for future work with different focuses.

The PPM Cancer dataset is the third dataset in this study. This study is the first study implementing process mining in this data. The complexity of the data served as an additional challenge in this study. Unlike the other two datasets, the PPM Cancer dataset is the complete database and a duplicate of the live database recording the clinical data of the patients in the PPM system at the LTHT. The issues of data quality and data understanding in this data were different. In the PPM Cancer dataset, data understanding was gained through frequent discussions with the development team and the clinical experts.

Apart from the patient records, the analysis in the PPM Cancer dataset was also supported by PPM Splunk recording real-time user access. The web-log provided in the PPM Splunk is also potentially useful for future work. The web-log in the PPM Splunk could be useful to reveal patient treatment and further analysis in the organisational perspective, such as to find if a specific clinician or a group of clinicians might have developed a best practice in doing their tasks.

The PPM EHR system was initially developed to support data collection of the national reporting on cancer treatment within the LTHT. This EHR system was then adopted to cover all services within the hospitals. It means that the data structure was changed from serving the needs of cancer reporting to supporting the day-to-day operations in the hospitals. Another aspect of the PPM Cancer dataset was that the PPM EHR system is an in-house software system. Discussions with the development team of the PPM revealed that the system had been through many changes over time. Apart from those changes, cancer treatment has also been changed over time.

7.6.2 The multi-level approach for identifying process change

The second contribution of this study was the multilevel approach for identifying process change. This approach resulted in a jointly-authored publication and was presented at the PODS4H 2019 in Vienna, Austria. The case study used in the paper was the analysis of the pathway from referral to the diagnosis of endometrial cancer, as described in Section 6.3. Some details of the approach were presented in Section 3.3.1.1. A summary and reflection of this approach are discussed in this section.

The general methods followed in the proposed approach are the L* Lifecycle Model and the Process Mining Project Methodology (PM^2); both have been described in Section 2.2.3.4. The general methodology was extended with a focus on the process analytics stage to analyse process changes over time. This approach allowed the detection of changes by comparing process execution from one year to another. Process change detection, localisation, and characterisation were performed at the model, trace, and activity levels. This was based on the understanding that a process can be represented as a process model, a set of traces, or a set of activity sequences. Those three levels have different levels of granularity. Comparison in those three levels revealed differences in three levels of granularity.

This approach was developed based on the available tools and plugins. The process discovery was done using the interactive Data-aware Heuristics Miner (iDHM) plugin in ProM. The conformance checking was done using the available plugins to calculate trace fitness, precision, and generalisation values. The process comparison was done on three different levels. The comparison in the process model level was performed using the same plugins as for the conformance checking. In the trace level, the duration and variant proportion were compared using bupaR in R. The comparison in the activity level was done based on the frequency and percentage of patients having a specific activity within a year.

This approach supports the exploration of process change analysis when the change is not known in advance. The graphical data visualisations were used to support discussions with the clinical experts about process evolutions. Future work could perhaps review the partitioning method, the comparison metrics, and the reference model discovery. The partitioning method carried out in the experiment was by partitioning the log into sub-logs based on the calendar year of diagnosis. Potential improvements could be to work on more detailed levels based on, for example, a monthly or weekly basis. The comparison metrics could be examined further to improve the process comparison. The reference model was discovered using iDHM, but other options could include using an inductive miner, fuzzy miner, or manually drawing the process model based on clinical guidance.

7.6.3 The signal decomposition approach for change analysis

The third contribution was the combination approach of signal decomposition and the Statistical Process Control (SPC) for process change analysis. This approach was applied to analyse a cohort of cancer patients in the PPM Cancer dataset, as presented in Section 6.5. This section provides a summary and reflection of this approach.

This approach was initially planned to improve the experiment of endometrial cancer pathways. The problem was that the number of patients in each sub-log based on the year of diagnosis was too small. It was a significant limitation because a small sample is not representative of the population. A small sample also breaches the information governance rule to exclude the granular information of fewer than six cases.

The input was the number of monthly records in each activity of interest. The central part of the approach was the signal decomposition that was done using an R library. The results are separated plots of the trend, seasonal, and random signals. It means that the number of monthly records was analysed to see the trend over time, the seasonal/monthly pattern, and the random/residual signal. The seasonal pattern was explained to see the months when the average number of records was at a minimum or a maximum value. The residual signal was analysed using SPC to detect change points outside the control lines.

This approach successfully revealed several types of system changes, as referred to as the dimensions of Leavitt’s diamond. Some technology changes were in 2003 when the PPM EHR system started to record cancer treatment events only and 2018 when *consultation* started to be recorded in another system. One organisational structure change was in 2010 when the PPM EHR system was underpinned to be used throughout the whole hospital. One task change was evidenced in 2014 when the haematology department started to deliver chemotherapy that previously could only be done by clinicians in the gynaecology and gynaecological oncology departments. This approach is potentially useful to be applied to other case studies with large datasets.

7.6.4 The time window selection to analyse process

The fourth and final contribution of this thesis was the time window selection to analyse the process. Process analysis studies generally work with a relatively long duration of data. The duration of the MIMIC-III dataset is 12 years (2001–2012), the PPM Chemotherapy dataset is nine years (2004–2012), and the PPM Cancer dataset is 16 years (2003–2018). The longer the duration, the more likely the data is prone to changes in the system and/or the process.

A common method in data analysis is to assume that the data is static, and the large volume of data is analysed at the same time. This study revealed the possibility that many changes had happened in the process, and those changes were evidenced in the data. The proposed multilevel approach for process change analysis could be used to analyse those changes. Some changes identified using this approach were the evolution of data collection within an EHR system, the increasing capacity of the service, the introduction of a new feature in the EHR system, and the additional assignment of a particular role.

One important issue in the process change analysis is the time window selection. In the PPM Cancer dataset case study, the event log was split into yearly sub-logs based on the calendar year of diagnosis. Further partitioning of the log into months, weeks, days, or hours could be considered in order to detect changes in a smaller window. The important points to consider are the expected duration of the process of interest and the size of the data within each time window. The window size should not be smaller than the expected process duration and the number of data points in each time window should not be smaller than the number required for an accurate analysis. For example, in the PPM Cancer dataset case study, the experiments were done to analyse the pathway from referral to diagnosis of endometrial cancer. The clinical experts suggested the inclusion of pathways of a maximum of 120 days. In this case, the window size should not be smaller than 120 days (4 months). If the window size is smaller than four months, many of the traces would be excluded in the study. The information governance rule of this study prevented the revealing of the information of six patients or less. This limitation made it not possible to make partitions of the event log into smaller sizes.

7.7 Summary

This chapter discussed the findings and reflections on the method and experiments of this research. Process change analysis as the main part of this research has been done using process mining. This approach has been applied to the three datasets. The challenges on working with healthcare data for process mining are related to many factors, including the data access and ethics approval, data quality, data understanding, data and process visualisations, and process change analysis. Despite those challenges, the method proposed in this research has been successfully applied to the three datasets to analyse process change over time. Contributions of this thesis include the three case studies for healthcare process mining, the multi-level approach for identifying process change, the signal decomposition approach for change analysis, and the time window selection to analyse the process.

Chapter 8
Summary

This thesis has presented a literature review, a methodology development, three case studies analysed using the methodology, and a discussion. The research objective was to use process mining approaches to detect and analyse process changes in the EHR systems. This chapter summarises the thesis by providing the conclusions and future work of this study.

8.1 Conclusions

This study has explored process mining methods for analysing data that were routinely collected within hospital information systems. Three case studies were presented: the MIMIC-III dataset, the Patient Pathway Manager (PPM) Chemotherapy dataset, and the PPM Cancer dataset. The conclusions of this research are summarised below following the thesis structure.

8.1.1 Conclusion from the literature review

A literature review of process mining in oncology was conducted in the early phase of this study. The important findings are as follows:

1) The *most common case study* in healthcare process mining is in cancer, more specifically in gynaecology. Process mining in gynaecology was found in 24 of the 37 papers (65%) reviewed in the early stages of this study. Even though cancer is the most common, a limited number of case studies are available. Previous studies have mostly used a dataset in the Business Process Intelligence Challenge (BPIC) 2011.
2) The *most common analysis* in healthcare process mining is from the process discovery and the control-flow perspective. Most studies worked in process discovery, as found in 35 of the 37 papers (95%) reviewed. Process discovery was commonly followed by a control-flow analysis to examine the relationship between one activity and the others.
3) The *most commonly used tool* for process mining in healthcare is the ProM framework. The ProM framework is a de facto standard in the process mining

research community and can be combined with other tools. ProM is also the primary tool used in this study, along with some additional tools, such as DISCO and bupaR.

4) *Limitations in healthcare process mining* projects include data, technique, and team limitations. Data limitation is related to the sensitivity and confidential nature of healthcare data. Technique limitation revealed an opportunity for method development. Team limitation is a reminder to include multidisciplinary experts in the team.

A further literature review was done during the research to explore publications in healthcare and technical backgrounds. Literature in the healthcare background includes healthcare systems in the US and the UK, EHR research, coding standards, process guidelines, and cancer. Literature in the technical background includes workflow technology, process modelling notations, process mining, process mining in healthcare, process mining for process change analysis, and statistical approach. One important point is that process change analysis is an understudied area in process mining projects. Most process mining projects worked with an assumption that the dataset is static. This research shows that this assumption is not reasonable for studies with long-duration datasets.

8.1.2 Conclusion from the methodology development

The main stages of the general methodology in this study were built based on the L* life-cycle model and the Process Mining Project Methodology (PM^2) as the two well-known process mining methods. Additional methods adopted in this study were the question-based methodology, the ClearPath method, concept drift analysis, signal decomposition, and SPC methods. The four main stages are: (1) planning and justification, (2) extraction, transformation, and loading (ETL), (3) mining and analysis, and (4) evaluation.

The main part of this study is the process analytics in the third stage. Process analytics in this study was done for process change analysis. The main steps in process change analysis are: change detection, change characterisation and localisation, and the unravelling of the process evolution. The general change detection was to split the event log into sub-logs based on time. Those sub-logs were then compared to see the variance over time. Change characterisation was done through a more in-depth analysis of the variability over time. The unravelling process evolution was done through discussions with clinical experts. There are two types of process change analysis based on the initial condition, which are with and without a known process change. Both types were explored in this research.

8.1.3 Conclusion from the experiments on the MIMIC-III dataset

The MIMIC-III dataset was the first dataset used in this study. It was suitable for this study because it is a publicly available healthcare dataset so that the study using this dataset is reproducible by other studies. There is no event log provided directly in the MIMIC-III database, but there are 16 event tables recording timestamped clinical events relating to patient treatments. Those tables can be easily transformed into an event log suitable for process mining. The database has been through a de-identification process and is available for research through ethical clearance. It is another advantage for a study to have a curated dataset that makes it ready for many types of analysis.

An assessment of the data quality found that, despite some limitations, the MIMIC-III dataset can be analysed with process mining and is useful for method development in the studies. For this study, the MIMIC-III dataset is specifically suitable to test the effect of system change. The documentation of the MIMIC-III dataset explained that the system from which the clinical data in MIMIC-III came was changed in 2008 from the CareVue (CV) system (2001–2008) to the MetaVision (MV) system (2008–2012).

The limitation of using the MIMIC-III dataset was due to its de-identification procedures. All dates recorded in the MIMIC-III database had been shifted, consistent for the same patient but randomly distributed in the future. It is mainly an issue in process mining that limits cross-patient analysis, such as workload and bottleneck analysis. Another obvious limitation was that there was no direct access to clinical experts from the hospital.

8.1.4 Conclusion from the experiments on PPM Chemotherapy dataset

The PPM Chemotherapy dataset was an extract from the PPM database with a specific focus on the clinical records of patients receiving chemotherapy treatment in the Leeds Teaching Hospitals NHS Trust (LTHT). This dataset had been used in a previous study and had been curated for that study.

In this study, the PPM Chemotherapy dataset was analysed by reproducing the previous study and improving it with a structured methodology for pathway analysis using process mining. The result was presented in a process model and dotted chart to show the flow of the activities during chemotherapy cycles. One important finding from the clinical perspective is that patients who underwent chemotherapy fell into three categories: those who stopped before the sixth cycle of chemotherapy, those who completed six cycles of chemotherapy without any complications, and those who had more complicated treatment over the years. The dotted chart presented those three groups of patients in a clear visualisation.

This dataset has also been used to analyse process change over time. It is evidenced that the conformance values of the process had been changed over the years. An additional analysis was done to explore the possibility of using a trace clustering approach to find different patterns within a process. For this cohort, there were changes in 2004–2005 and 2007–2008. The conformance-based comparison was the only comparison performed on the PPM Chemotherapy dataset and is a primary method to compare processes over time.

8.1.5 Conclusion from the experiments on the PPM Cancer dataset

The third case study was done using the PPM Cancer dataset. Among the three datasets used in this study, the PPM Cancer dataset was the most complex. Access was given to the complete database, which is a copy of the original raw version of the live database in the PPM Electronic Health Record (EHR) system. The advantage of using this dataset was that it represented the real data in a large hospital in the UK. The PPM dataset also came with a direct connection to the clinical experts and the development team within the hospital.

Four experiments were conducted in this case study. The first experiment was done to analyse a known change in the EHR system and its effect on the cancer treatment. The change of interest is the GP tab introduction, which was a part of the Leeds Care Record (LCR) initiative to integrated care records from many providers including GP and hospital. In this experiment, the GP tab introduction was related to the chemotherapy cycles of breast cancer patients. The records of user access in the PPM Splunk was combined with the records of patient treatment in the PPM Cancer dataset. The findings showed potentially useful insights to examine the effects of the introduction of a feature in the EHR system to the cancer treatment. The second and third experiments were done to analyse the pathways of endometrial cancer patients. Those two experiments resulted in two methods to investigate process change over time, when there was no prior information about any change in the system. The fourth experiment was done to explore all events related to cancer treatment in the PPM database for process change analysis.

8.1.6 Conclusion from the discussion

This study is a process-oriented data analytics study using process mining and process change analysis approaches. The datasets analysed in this study were the routinely collected clinical data from the EHR system. The advantages were that those datasets recorded the real execution of treatment processes within the hospital and that no further effort was required for data creation. The limitations were that the data quality was dependent on many factors, including the data collection, data recording, and the data management of the EHR system. A specific characteristic of the datasets is the high veracity of the data, which leads to high variability over time due to many possible changes that happened in the system.

The challenges on the healthcare process mining included data access and ethics approval, data quality, data understanding, data and process visualisation, and process change over time. The process change over time was the main challenge explored in this study. The dimensions of process change analysis are the modes of handling, change duration, perspective, the pattern of change, sub-problems, and the nature of change. One crucial aspect is the time window selection to analyse the process. This aspect has been explored in different experiments in this research, especially using the PPM Cancer dataset with the monthly and yearly record analysis.

8.2 Presentation and feedback

The contributions of this thesis as described in Section 7.5.4 are the case studies for healthcare process mining, the multi-level approach for identifying process change, the signal decomposition approach for change analysis, and the time window selection to analyse the process. These contributions can be seen as a way to contribute to the community of healthcare process mining, more specifically to the alliance of Process-Oriented Data Science for Healthcare (PODS4H), and in the broader communities of related studies. This study contributes to the technical and clinical aspects of the healthcare process mining studies.

Published contributions of this study included three posters presented at international conferences, four paper presented at international conferences, and one paper published in an international journal, as listed in the front pages of this study. Other than those eight publications, some parts of this study have been presented in seven further events to gather feedback from a range of communities. Those events are:

1) Guest lecture in the MSc class of data mining (5 May 2016)

This was an hour-long joint-presentation with Owen Johnson on an introduction to process mining. There were around 30 students who attended the lecture. I presented an illustrative example of analysing healthcare data using process mining. The dataset for this presentation was a fictional data from Leeds Accident and Emergency (A&E). The students were actively trying to find patterns in the small event log. This presentation is not explicitly presented in this thesis but has developed the foundation of understanding process mining as the fundamental approach in this study.

2) LIDA Seminar: health informatics and data analytics (10 November 2016)

This is one of the Leeds Institute of Data Analytics (LIDA) seminar series. I presented the idea of this study to test the effect of UI design on the healthcare process. Other presenters in this seminar were Owen Johnson and Professor John Fox. Owen presented a general introduction to mining, modelling, and improving care pathways with big data. I presented my work as a case study of process mining in oncology. Professor John Fox is the main speaker presented Artificial Intelligence (AI) in medicine: data science meets knowledge management. He is a professor at Oxford University and the Chairman and Co-founder of OpenClinical.

One insight gained from the presentations was that organisations are complex systems where task, technology, people, and structure are interrelated and mutually adjusting. One important piece of feedback given by a data scientist after the seminar was that healthcare data are naturally messy, and that might provide a real challenge for my study. This has been addressed by applying a quality assessment of the datasets.

3) School of Computing PhD symposium (17 January 2017)

This is an annual PhD symposium within the School of Computing, University of Leeds. There were around 20 people attended the presentation. I introduced myself and my study plan for this thesis. The material presented was based on the literature review of process mining in oncology, as published in the conference in 2016.

An interesting piece of feedback from this symposium was that I would need to define the characteristics of the UI change and not confuse it with other types of changes. This has been reflected in the fourth research question in this study, where a process change is characterised by a significant change in one or more metrics of the multi-level approach.

4) *WUN Data Science Thematic Workshop (1 May 2017)*

This was the World University Network (WUN) data science thematic workshop: wellness data for healthy societies. The workshop was presented in New York, as part of the annual WUN Conference and AGM. Eric Rojas (Pontifical Catholic University of Chile) and I joined in Owen Johnson (the main supervisor of this research) presentation. I presented the method and progress of our studies in the comparative analytics of patient pathways data.

One interesting insight from this workshop was that there were several groups in other countries working on pathway analysis closely related to my study. Their concern during the workshop was on finding a general method to be applied to different countries. A follow-up discussion occurred in Cambridge to explore the possibility of joint research between Leeds and Cambridge hospitals. The publication on the data quality assessment of the MIMIC-III is written in collaboration with Eric Rojas as a follow-up result of this workshop [168].

5) *PHE NCRAS seminar (23 May 2018)*

This is one of the routine seminars hosted by Public Health England National Cancer Registration and Analysis Service (PHE NCRAS). This event was held in London and was broadcast throughout the PHE network. It was a joint-presentation with Owen Johnson entitled “Process mining of cancer pathways using Electronic Health Records”. Owen presented some challenges understanding real pathways of care and how EHR can help. I presented some of the progress made in the analysis of the PPM Chemotherapy and the PPM Full datasets.

One interesting insight from this seminar was that PHE NCRAS has a team dedicated to pathway analysis, but they have not used any process mining approach in their work. Further exploration on the official website of PHE NCRAS found information about cancer treatment, cancer statistics, routes to diagnosis, cancer outcome metrics, and other topics related to this study.

6) School of Computing PhD Symposium (14 February 2019)

This was an annual PhD symposium within the School of Computing, University of Leeds. I presented the method and progress of my study in this thesis. The main focus of the presentation was on the analysis of process change over time. The material presented in this event was the interim result as later published in the PODS4H 2019.

One interesting discussion was about how best to present the results of this study to the clinical experts. This discussion is addressed in Section 7.3.4 on data and process visualisation. The outstanding discussion is that there is a range of options to visualise the results. The choice should be taken by considering what the most important message to relay to the clinical experts is.

7) Manchester healthcare analytics group seminar (14 June 2019)

This was an internal seminar held by the healthcare analytics group, University of Manchester. Eric Rojas (Chile), Frank Fox (Ireland; a member of our healthcare process mining research group), and I presented our studies in healthcare process mining. I presented some results from my experiments with the PPM Chemotherapy and the PPM Cancer datasets.

One interesting discussion in this seminar was about the meaningfulness of the results from the clinical perspective. This can be seen as a potential to use the question-driven approach, and this has been addressed in the general method of this research. An engagement with the clinical experts is needed during the research.

8.3 Future work

This study shows some potentials to be used and improved in future work in healthcare process mining studies. Future work could improve this study in five directions, as described in the following paragraphs.

The first direction could be to improve the method to capture patient characteristics in the analysis. This improvement would enhance the usefulness of the study for the clinical experts. This is also based on the feedback that this study could be more useful for epidemiology study. By improving the method to capture patient characteristics and relating those characteristics to the process mining results, process mining can be more useful in an epidemiology study.

The second direction is the potential to use process mining to create event log by combining software usage log with patient treatment records. Experiment 5 with the GP tab analysis shown that this approach is potentially useful to add another perspective in the typical process mining approach. Software usage log could be included in the process discovery step to enhance the process model representing the patient treatment pathways with other actions done by the users during patient treatment.

The third direction could be to support the interpretation of the results by statistical analysis. Regular complaints by clinical and statistical experts are that this research lack of statistical evidence. Interpreting process mining results using statistical analysis could demonstrate stronger evidence of the significance of the results. Some statistical approaches have been applied in this study, but there are still opportunities to improve this analysis with a more sophisticated statistical approach.

The fourth direction could be to widen the use of the proposed methods in healthcare case studies. This could include many different cohorts of patients and the exploration of many other guidelines and expected pathways within the guidelines. This thesis provides evidence that the general method can be applied in two different data sources and three different datasets. The multi-level approach to detect process change over time is relatable to many real-life settings in the hospitals. Understanding the change over time provides a valuable insight to get a better understanding of the dynamic nature of the healthcare processes.

The fifth and the last direction could be to improve the method for international comparison of healthcare processes. The datasets included in this study were from the USA, a country with a non-universal healthcare system dominated by private providers, and from the UK, a country with a universal government-funded healthcare system. Other countries have healthcare systems similar to either one of these or a combination of both. The successful implementation of this method in the datasets from the USA and the UK open up an opportunity for international comparison of healthcare processes.

8.4 Final remark

Process mining has been applied to many studies in the healthcare domain. One understudied area is the analysis of process change. When process mining is applied to a dataset over the years, the process might have changed, and the change might evidence in the data. This thesis explores the opportunity to use process mining approach to analyse process change over time. The method proposed in this study has been applied to three different datasets. It is shown that process mining can be used to analyse process change over time. When a change is known in the initial stage of the study, process mining can be used to analyse the process before and after the change point. When a change is unknown, process mining can be used to detect change points by creating partitions of the data over time and comparing the process characteristics in the subsequent partitions.

On the other side, EHR systems are generally evolving. The EHR systems might be changed because of four factors affecting process and data, which are the structure, process, technology, and people. This thesis focused on the impact of UI change on clinical pathways, but those four factors of change are inseparable in the discussions during the research. Process mining studies need to consider the interplay between those four factors when analysing the impact of changes in the EHR systems.

List of References

[1] R. Lenz and M. Reichert, "IT support for healthcare processes – premises, challenges, perspectives," *Data & Knowledge Engineering*, vol. 61, no. 1, pp. 39–58, 2007.

[2] R. Lenz, T. Elstner, H. Siegele, and K. a Kuhn, "A practical approach to process support in health information systems.," *Journal of the American Medical Informatics Association : JAMIA*, vol. 9, no. 6, pp. 571–585, 2002.

[3] W. M. P. van der Aalst, A. Adriansyah, A. K. A. de Medeiros, F. Arcieri, T. Baier, T. Blickle, J. C. Bose, P. van der Brand, R. Brandtjen, J. Buijs, A. Burattin, J. Carmona, M. Castellanos, J. Claes, and J. Cook, "Process Mining Manifesto," *Business Process Management Workshops*, vol. 99, pp. 169–194, 2011.

[4] W. M. P. van der Aalst, *Process Mining: Data Science in Action*, 2nd ed. Springer-Verlag Berlin Heidelberg, 2016.

[5] A. K. Jha, C. M. Desroches, E. G. Campbell, K. Donelan, S. R. Rao, T. G. Ferris, A. Shields, S. Rosenbaum, and D. Blumenthal, "Use of electronic health records in U.S. Hospitals," *New England Journal of Medicine*, vol. 360, no. 16, pp. 1628–1638, 2009.

[6] K. M. Cresswell, A. Worth, and A. Sheikh, "Integration of a nationally procured electronic health record system into user work practices," *BMC Medical Informatics and Decision Making*, vol. 12, no. 1, pp. 1–12, 2012.

[7] P. Gooch and A. Roudsari, "Computerization of workflows, guidelines, and care pathways: a review of implementation challenges for process-oriented health information systems.," *Journal of the American Medical Informatics Association : JAMIA*, vol. 18, no. 6, pp. 738–48, 2011.

[8] M. Schnabel, M. Bäumlein, R. Lenz, C. Biber, R. Blaser, O. Heger, and M. Beyer, "IT support for clinical pathways—Lessons learned," *International Journal of Medical Informatics*, vol. 76, pp. S397–S402, 2007.

[9] I. Atastina and A. P. Kurniati, "Student registration process evaluation using process mining case study: IT Telkom," in *2014 9th International Conference on Digital Information Management, ICDIM 2014*, 2014, pp. 189–193.

[10] S. Dadashnia, T. Niesen, P. Hake, P. Fettke, N. Mehdiyev, and J. Evermann, "Identification of Distinct Usage Patterns and Prediction of Customer Behavior," *Sixth International Business Process Intelligence Challenge (BPIC-2016)*, no. September, 2016.

[11] R. S. Mans, W. M. P. van der Aalst, and R. J. B. Vanwersch, "Process Mining in Healthcare," vol. 42, pp. 9236–9251, 2015.

[12] National Health System, "Delivering Cancer Waiting Times: A Good Practice Guide," pp. 0–67, 2015.

[13] B. Shneiderman, C. Plaisant, M. (Maxine S. . Cohen, S. M. Jacobs, and N. Elmqvist, *Designing the user interface : strategies for effective human-computer interaction.* .

[14] J. Carmona, C. Fernandez-Llatas, R. Gatta, O. Johnson, N. Martin, J. Munoz-Gama, E. Rojas, L. Sacchi, F. Seoane, M. Sepulveda, and V. Traver, "Alliance – PODS4H," 2019. .

[15] 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*, 2015, pp. 297–313.

[16] R. P. J. C. Bose, W. M. P. Van Der Aalst, I. Žliobaite, and M. Pechenizkiy, "Handling concept drift in process mining," in *International Conference on Advanced Information Systems Engineering*, 2011, pp. 391–405.

[17] S. Jonas, R. L. Goldsteen, and K. Goldsteen, *An introduction to the U.S. health care system*. Springer, 2007.

[18] K. Grosios, P. B. Gahan, and J. Burbidge, "Overview of healthcare in the UK," *EPMA Journal*, vol. 1, no. 4. pp. 529–534, 2010.

[19] Beth Israel Deaconess Medical Center, "Beth Israel Deaconess Medical Center," 2012. [Online]. Available: http://www.bidmc.org/. [Accessed: 20-Mar-2019].

[20] U.S. Census Bureau, "Population and Housing Unit Estimates," *Annual Estimates of the Resident Population for the United States, Regions, States, and Puerto Rico*, 2018. [Online]. Available: https://www.census.gov/programs-surveys/popest/data/tables.2018.html. [Accessed: 30-Jul-2019].

[21] Leeds Teaching Hospitals NHS Trust, "Leeds Teaching Hospital," 2016. [Online]. Available: http://www.leedsth.nhs.uk/. [Accessed: 26-Jul-2016].

[22] Office for National Statistics, "Estimates of the population for the UK, England and Wales, Scotland and Northern Ireland," 2018. [Online]. Available: https://www.ons.gov.uk/peoplepopulationandcommunity/populationandmigration/populationestimates/datasets/populationestimatesforukenglandandwalesscotlandandnorthernireland. [Accessed: 30-Jul-2019].

[23] W. Hazell, "Analysed: The biggest NHS providers of specialised services | News | Health Service Journal," 2015. [Online]. Available: https://www.hsj.co.uk/home/analysed-the-biggest-nhs-providers-of-specialised-services/5091147.article. [Accessed: 30-Jul-2019].

[24] O. A. Johnson and S. E. Abiodun, "Understanding What Success In Health Information Systems Looks Like: The Patient Pathway Management (PPM) System At Leeds," *UK Academy for Information Systems Conference Proceedings*, no. 22, 2011.

[25] A. P. Kurniati, O. Johnson, D. Hogg, and G. Hall, "Process Mining in Oncology: a Literature Review," in *The 6th ICICM, IEEE*, 2016.

[26] World Health Organization, "Health systems," 2019. [Online]. Available: https://www.who.int/topics/health_systems/en/.

[27] World Health Organization, "Strengthening health systems to improve health outcomes," 2007.

[28] Department for Professional Employees, "The US health care system: an international perspective," 2016.

[29] B. T. Powell, "The structure of the NHS in England," *Private complaints and public health*, no. July, pp. 85–92, 2017.

[30] K. Häyrinen, K. Saranto, and P. Nykänen, "Definition, structure, content, use and impacts of electronic health records: A review of the research literature," *International Journal of Medical Informatics*, vol. 77, no. 5, pp. 291–304, 2008.

[31] R. H. Fletcher, S. W. Fletcher, and G. S. Fletcher, *Clinical epidemiology: The essentials*, 5th ed. 2012.

[32] S. Janmohamed, S. Goldman, S. Ong, F. Fritz, S. Duclaux, A. Zalewski, J. P. Pell, F. Zannad, A. Michel, M. Thoenes, M. R. Southworth, I. Ford, J. I. Blomster, M. Leenay, W. G. Stough, L. H. Curtis, J. Kreuzer, and M. R. Cowie, "Electronic health records to facilitate clinical research," *Clinical Research in Cardiology*, vol. 106, no. 1, pp. 1–9, 2016.

[33] Z. Huang, X. Lu, and H. Duan, "On mining clinical pathway patterns from medical behaviors," *Artificial Intelligence in Medicine*, vol. 56, no. 1, pp. 35–50, 2012.

[34] K. Baker, E. Dunwoodie, R. G. Jones, A. Newsham, O. Johnson, C. P. Price, J. Wolstenholme, J. Leal, P. McGinley, C. Twelves, and G. Hall, "Process mining routinely collected electronic health records to define real-life clinical pathways during chemotherapy," *International Journal of Medical Informatics*, vol. 103, pp. 32–41, 2017.

[35] World Health Organization, *International Classification of Diseases*, vol. 10th rev.

World Health Organization, 2004.

[36] World Health Organization, “International Classification of Diseases, 11th Revision (ICD-11),” 2018.

[37] National Institute for Health and Care Excellence, “NATIONAL INSTITUTE FOR HEALTH AND CARE Business Plan : objectives and performance measures,” no. April 2018, 2018.

[38] National Institute for Health and Care Excellence, “Advanced breast cancer overview,” 2019. [Online]. Available: https://pathways.nice.org.uk/pathways/advanced-breast-cancer.

[39] Leeds Teaching Hospitals NHS Trust, “Leeds Health Pathways,” 2019. [Online]. Available: http://nww.lhp.leedsth.nhs.uk/.

[40] World Health Organization, “Cancer Fact Sheet,” 2018. [Online]. Available: http://www.who.int/mediacentre/factsheets/fs297/en/. [Accessed: 07-Jan-2019].

[41] CRUK, “Cancer Statistics for the UK,” *Cancer Research UK*, 2018. [Online]. Available: https://www.cancerresearchuk.org/health-professional/cancer-statistics-for-the-uk. [Accessed: 07-Jan-2019].

[42] D. Mort, “For Better, for Worse?: A Review of the Care of Patients who Died Within 30 Days of Receiving Systemic Anti-cancer Therapy,” 2008.

[43] National Cancer Institute, “Colorectal Cancer - Patient Version,” *national institutes of health*, 2017. [Online]. Available: https://www.cancer.gov/types/colorectal/patient/colon-treatment-pdq. [Accessed: 10-Jan-2019].

[44] National Institute for Health and Care Excellence, “Suspected cancer: recognition and referral,” 2015.

[45] J. Maddams, M. Utley, and H. Møller, “Projections of cancer prevalence in the United Kingdom, 2010-2040.,” *British journal of cancer*, vol. 107, no. 7, pp. 1195–202, 2012.

[46] National Institute for Health and Care Excellence, “Colorectal cancer: diagnosis and management,” *Clinical guideline*, 2011. .

[47] PDQ Adult Treatment Editorial Board, “PDQ Colon Cancer Treatment,” Bethesda.

[48] National Cancer Institute, “Breast Cancer - Patient Version,” 2018. [Online]. Available: https://www.cancer.gov/types/breast/patient/breast-treatment-pdq.

[49] National Institute for Health and Care Excellence, “Advanced breast cancer: diagnosis and treatment,” *Clinical guideline*, 2009. [Online]. Available: https://www.nice.org.uk/guidance/cg81.

[50] National Institute for Health and Care Excellence, “Familial breast cancer: classification, care and managing breast cancer and related risks in people with a family history of breast cancer,” 2013. [Online]. Available: https://www.nice.org.uk/guidance/cg164.

[51] Leeds Teaching Hospitals NHS Trust, “2WW Ideal Breast Pathway,” 2016. [Online]. Available: http://nww.lhp.leedsth.nhs.uk/referral_info/detail.aspx?ID=383.

[52] National Cancer Institute, “Uterine Cancer - Patient Version,” 2018. [Online]. Available: https://www.cancer.gov/types/uterine/patient/endometrial-treatment-pdq. [Accessed: 20-Aug-2007].

[53] National Institute for Health and Care Excellence, “Laparoscopic hysterectomy (including laparoscopic total hysterectomy and laparoscopically assisted vaginal hysterectomy) for endometrial cancer,” *Guidance*, 2010. [Online]. Available: https://www.nice.org.uk/Guidance/IPG356.

[54] Leeds Teaching Hospitals NHS Trust, “2WW Endometrial Pathway,” 2015. [Online]. Available: http://nww.lhp.leedsth.nhs.uk/referral_info/detail.aspx?ID=371.

[55] NHS, “Waiting times for suspected and diagnosed cancer patients: 2018-2019 annual report,” 2019.

[56] D. Hollingsworth, “The Workflow Management Coalition -The Workflow Reference Model,” Document Number TC00-1003 19, 1993.

[57] B. Kiepuszewski, A. P. Barros, and A. Dogac, “Workflow Patterns,” *Distributed and Parallel Databases*, vol. 14, no. 1, pp. 5–51, 2003.

[58] M. Beaudouin-Lafon, *Computer Supported Co-operative Work (CSCW)*. New Jersey, USA: John Wiley & Sons, 1999.

[59] H. R. Lewis and C. H. Papadimitriou, *Elements of the theory of computation*, 2nd ed. Prentice Hall PTR, 1997.

[60] M. Nüttgens, T. Feld, and V. Zimmermann, “Business Process Modeling with EPC and UML: Transformation or Integration?,” in *The Unified Modeling Language*, 2012, pp. 250–261.

[61] H. Eriksson and M. Penker, “Business Modeling With UML,” *Business Patterns at Work*, p. 12, 2000.

[62] OMG, “Business Process Modeling Notation (BPMN) Version 1.0,” *OMG Final Adopted Specification, Object Management Group 190*, 2006. .

[63] C. A. Petri, “Communication with Automata (PhD dissertation),” Technischen Hochschule Darmstadt, 1962.

[64] R. M. Keller, “Formal verification of parallel programs,” *Communications of the ACM*, vol. 19, no. 7, pp. 371–384, 1976.

[65] M. Kot, “The State Explosion Problem,” *Lectures on Petri Nets I: Basic Models*, vol. 1491, no. series Lecture Notes in Computer Science, pp. 429–528, 2008.

[66] N. Russell, W. M. P. van der Aalst, A. H. M. ter Hofstede, and P. Wohed, “On the Suitability of UML 2.0 Activity Diagrams for Business Process Modelling,” in *Conceptual Modelling 2006: Proceedings of APCCM2006*, 2006, no. January, pp. 16–19.

[67] P. Wohed, W. M. P. van der Aalst, M. Dumas, A. H. M. ter Hofstede, and N. Russell, “On the suitability of BPMN for Business Process Modelling,” in *International Conference on Business Process Management*, 2006, pp. 161–176.

[68] R. Müller and A. Rogge-Solti, “BPMN for healthcare processes,” *CEUR Workshop Proceedings*, vol. 705, no. May 2014, pp. 65–72, 2011.

[69] H. Scheuerlein, F. Rauchfuss, Y. Dittmar, R. Molle, T. Lehmann, N. Pienkos, and U. Settmacher, “New methods for clinical pathways-Business Process Modeling Notation (BPMN) and Tangible Business Process Modeling (t.BPM).,” *Langenbeck’s archives of surgery / Deutsche Gesellschaft für Chirurgie*, vol. 397, no. 5, pp. 755–61, 2012.

[70] C. A. Ellis and G. J. Nutt, “Modeling and enactment of workflow systems,” in *International Conference on Application and Theory of*, 2012, pp. 1–16.

[71] W. M. P. van der Aalst, “Three Good reasons for Using a Petri-net-based Workflow Management System,” *Information and Process Integration in Enterprises: Rethinking Documents*, pp. 161–182, 1998.

[72] W. M. P. van der Aalst and C. Stahl, *Modeling business processes: a Petri net-oriented approach*. MIT Press, 2011.

[73] M. Lang, T. Bürkle, S. Laumann, and H.-U. Prokosch, “Process Mining for Clinical Workflows: Challenges and Current Limitations,” *eHealth Beyond the Horizon - Get IT There (Proceedings 21st International Congress of the European Federation for Medical Informatics, MIE)*, pp. 229–234, 2008.

[74] W. M. P. van der Aalst, *Process mining: discovery, conformance and enhancement of business processes*, 1st ed. Springer-Verlag Berlin Heidelberg, 2011.

[75] W. M. P. van der Aalst, “Workflow mining: Discovering process models from event logs,” *IEEE Transactions on Knowledge and Data Engineering*, vol. 16, no. 9, pp. 1128–1142, 2004.

[76] A. K. A. de Medeiros, W. M. P. van der Aalst, and A. J. M. M. Weijters, “Workflow Mining: Current Status and Future Directions,” no. i, pp. 389–406, 2010.

[77] C. Günther and W. M. P. van der Aalst, “Fuzzy mining–adaptive process simplification based on multi-perspective metrics,” *Business Process Management*, 2007.

[78] S. J. J. Leemans, D. Fahland, and W. M. P. 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*, 2013, pp. 311–329.

[79] S. J. J. Leemans, D. Fahland, and W. M. P. van der Aalst, “Discovering block-structured process models from event logs containing infrequent behaviour,” in *International conference on business process management*, 2013, pp. 66–78.

[80] S. J. J. Leemans, D. Fahland, and W. M. P. Van Der Aalst, “Discovering block-structured process models from incomplete event logs,” *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 8489 LNCS, pp. 91–110, 2014.

[81] A. J. M. M. Weijters, W. M. P. van der Aalst, and A. A. De Medeiros, “Process mining with the heuristics miner-algorithm,” *Technische Universiteit Eindhoven, Tech. Rep. WP 166*, vol. 166, pp. 1–34, 2006.

[82] X. Zhang and S. Chen, “Pathway identification via process mining for patients with multiple conditions,” in *IEEE International Conference on Industrial Engineering and Engineering Management*, 2012, pp. 1754–1758.

[83] F. Mannhardt, M. De Leoni, and H. A. Reijers, “Heuristic mining revamped: An interactive, data-Aware, and conformance-Aware miner,” *CEUR Workshop Proceedings*, vol. 1920, no. August 2010, pp. 358–359, 2017.

[84] G. Janssenswillen, N. Donders, T. Jouck, and B. Depaire, “A comparative study of existing quality measures for process discovery,” *Information Systems*, vol. 71, pp. 1–15, 2017.

[85] W. M. P. van der Aalst, A. Adriansyah, and B. van Dongen, “Replaying history on process models for conformance checking and performance analysis,” *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, vol. 2, no. 2, pp. 182–192, 2012.

[86] 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,” in *OTM Confederated International Conferences*, 2012, pp. 305–322.

[87] M. Bozkaya, J. Gabriels, J. Werf, B. M., G. J., and V. D. W. J.M., “Process diagnostics: A method based on process mining,” in *International Conference on Information, Process, and Knowledge Management, eKNOW 2009*, 2009, no. 1, pp. 22–27.

[88] A. Rebuge and D. R. D. Ferreira, “Business process analysis in healthcare environments: A methodology based on process mining,” *Information Systems*, vol. 37, pp. 99–116, 2012.

[89] E. Rojas, M. Sepúlveda, J. Munoz-Gama, D. Capurro, V. Traver, and C. Fernandez-Llatas, “Question-Driven Methodology for Analyzing Emergency Room Processes Using Process Mining,” *Applied Sciences*, vol. 7, no. 3, p. 302, 2017.

[90] W. Zhou and S. Piramuthu, “Framework, strategy and evaluation of health care processes with RFID,” *Decision Support Systems*, vol. 50, no. 1, pp. 222–233, Dec. 2010.

[91] O. A. Johnson, T. B. A. Dhafari, A. Kurniati, and E. Rojas, “The ClearPath Method for Care Pathway Process Mining and Simulation,” in *Lecture Notes in Business Information Processing*, 2018, pp. 1–12.

[92] B. S. Meeting and P. Chapman, “The CRISP-DM User Guide,” *The CRISP-DM User Guide*, p. 14, 1999.

[93] C. W. Günther and A. Rozinat, “Disco: Discover Your Processes.,” in *BPM 2012 Demonstration Track*, 2012, vol. 940, pp. 40–44.

[94] 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 *International Conference on Application and Theory of Petri Nets*, 2005, pp. 444–454.

[95] G. Janssenswillen, “bupaR: Business Process Analysis in R,” *R package version 0.4.2*, 2019. [Online]. Available: https://cran.r-project.org/package=bupaR.

[96] R. S. Mans, M. H. Schonenberg, M. Song, W. M. P. van der Aalst, and P. J. M. Bakker, “Application of Process Mining in Healthcare – A Case Study in a Dutch Hospital,” *Proceedings of BIOSTEC 2008*, vol. 25, pp. 425–438, 2008.

[97] R. Mans, H. Reijers, M. van Genuchten, and D. Wismeijer, “Mining processes in dentistry,” *Proceedings of the 2nd ACM SIGHIT symposium on International health informatics - IHI ’12*, p. 379, 2012.

[98] J. B. van Osch Dekker, “Process Mining: Acquiring Objective Process Information for Healthcare Process Management with the CRISP-OM Framework I PREFACE,” Eindhoven University of Technology, 2008.

[99] R. Mans, M. Schonenberg, M. Song, W. van der Aalst, and P. Bakker, “PROCESS MINING IN HEALTHCARE A Case Study,” in *Healthinf 2008*, 2008, pp. 118–125.

[100] W. Yang and Q. Su, “Process Mining for Clinical Pathway Literature Review and Future Directions,” *Service Systems and Service Management (ICSSSM), 2014 11th International Conference*, pp. 1–5, 2014.

[101] E. Rojas and J. Munoz-Gama, “Process mining in healthcare: A literature review,” *Journal of biomedical informatics*, vol. 61, pp. 224–236, 2016.

[102] M. Song, C. W. Gunther, and W. M. P. van der Aalst, “Trace clustering in process mining,” in *Lecture Notes in Business Information Processing*, 2009.

[103] B. W. Boudewijn van Dongen, Diogo R. Ferreira, “Business Process Intelligence Challenge (BPIC),” *IEEE Task Force on Process Mining*, 2011. .

[104] R. P. J. C. Bose, R. S. Mans, and W. M. P. van der Aalst, “Wanna improve process mining results ? It ’ s high time we consider data quality issues seriously,” *BPM reports*, vol. 1302, 2013.

[105] E. Ramezani, D. Fahland, and W. M. P. van der Aalst, “Supporting domain experts to select and configure precise compliance rules,” *International Conference on Business Process Management. Springer International Publishing*, pp. 498–512, 2013.

[106] A. Deokar and J. Tao, “Semantics-based event log aggregation for process mining and analytics,” *Information Systems Frontiers*, vol. 17, pp. 1209–1226, 2015.

[107] E. Bellodi, F. Riguzzi, and E. Lamma, “Statistical Relational Learning for Workflow Mining,” *Intelligent Data Analysis*, vol. 20, no. 3, pp. 515–541, 2015.

[108] D. Antonelli, E. Baralis, G. Bruno, S. Chiusano, N. A. Mahoto, and C. Petrigni, “Analysis of diagnostic pathways for colon cancer,” *Flexible Services and Manufacturing Journal*, vol. 24, no. 4, pp. 379–399, 2012.

[109] H. Duan, L. Ji, X. Lu, Z. Huang, W. Dong, and C. Gan, “Discovery of clinical pathway patterns from event logs using probabilistic topic models,” *Journal of Biomedical Informatics*, vol. 47, pp. 39–57, 2013.

[110] J. De Weerdt, F. Caron, J. Vanthienen, and B. Baesens, “Getting a Grasp on Clinical Pathway Data: An Approach Based on Process Mining,” *Emerging Trends in Knowledge Discovery and Data Mining*, pp. 22–35, 2012.

[111] M. Räim, C. Di Ciccio, F. F. M. Maggi, M. Raeim, C. Di Ciccio, F. F. M. Maggi, M. Mecella, and J. Mendling, “Log-Based Understanding of Business Processes through Temporal Logic Query Checking,” *On the Move to Meaningful Internet Systems: Otm 2014 Conferences*, vol. 8841, no. October 2014, pp. 75–92, 2014.

[112] H. Syed and A. Das, “Temporal Needleman-Wunsch,” *IEEE International Conference on Data Science and Advanced Analytics (DSAA), 2015. 36678 2015.*, pp. 1–9, 2015.

[113] F. Ju, H. K. Lee, R. U. Osarogiagbon, X. Yu, N. Faris, and J. Li, "Computer modeling of lung cancer diagnosis-to-treatment process.," *Translational lung cancer research*, vol. 4, no. 4, pp. 404–14, 2015.

[114] A. Leontjeva and R. Conforti, "Complex Symbolic Sequence Encodings for Predictive Monitoring of Business Processes," *International Conference on Business Process Management. Springer International Publishing*, pp. 297–313, 2015.

[115] Z. Huang, X. Lu, and H. Duan, "Latent treatment pattern discovery for clinical processes," *Journal of medical systems*, 2013.

[116] P. Delias, M. Doumpos, E. Grigoroudis, P. Manolitzas, and N. Matsatsinis, "Supporting healthcare management decisions via robust clustering of event logs," *Knowledge-Based Systems*, vol. 84, pp. 203–213, 2015.

[117] J. J. Boere, "An Analysis and Redesign of the ICU Weaning Process using Data Analysis and Process Mining," Maastricht University Medical Centre, 2013.

[118] A. Partington, M. Wynn, S. Suriadi, C. Ouyang, and J. Karnon, "Process Mining for Clinical Processes: A comparative analysis of four Australian hospitals," *ACM Transactions on Management Information Systems*, 2015.

[119] M. Binder, W. Dorda, G. Duftschmid, R. Dunkl, K. A. Fröschl, W. Gall, W. Grossmann, K. Harmankaya, M. Hronsky, S. Rinderle-Ma, C. Rinner, and S. Weber, "On analyzing process compliance in skin cancer treatment: An experience report from the evidence-based medical compliance cluster (EBMC 2)," in *International Conference on Advanced Information Systems Engineering*, 2012, pp. 398–413.

[120] C. Di Francescomarino and M. Dumas, "Clustering-Based Predictive Process Monitoring," *arXiv preprint arXiv:1506.01428*, 2015.

[121] J. Meier, A. Dietz, A. Boehm, and T. Neumuth, "Predicting treatment process steps from events," *Journal of Biomedical Informatics*, vol. 53, pp. 308–319, 2015.

[122] F. Maggi, R. Bose, and W. M. P. van der Aalst, "A knowledge-based integrated approach for discovering and repairing declare maps," *International Conference on Advanced Information Systems Engineering. Springer Berlin Heidelberg*, pp. 433–448, 2013.

[123] F. M. Maggi, C. Di Francescomarino, M. Dumas, and C. Ghidini, "Predictive monitoring of business processes," *International Conference on Advanced Information Systems Engineering. Springer International Publishing*, pp. 457–472, 2014.

[124] F. M. Maggi, "Discovering metric temporal business constraints from event logs," *Lecture Notes in Business Information Processing*, vol. 194, pp. 261–275, 2014.

[125] X. Lu and R. Mans, "Conformance checking in healthcare based on partially ordered event data," *Proceedings of the 2014 IEEE Emerging Technology and Factory Automation (ETFA)*, pp. 1–8, 2014.

[126] U. Kaymak, R. Mans, T. Van De Steeg, and M. Dierks, "On Process Mining in Health Care," *IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, pp. 1859–1864, 2012.

[127] A. Burattin and M. Cimitile, "Online Discovery of Declarative Process Models from Event Streams," *IEEE Transactions on Services Computing*, vol. 8, no. 6, pp. 833–846, 2015.

[128] J. Poelmans, G. Dedene, G. Verheyden, H. Van Der Mussele, S. Viaene, and E. Peters, "Combining business process and data discovery techniques for analyzing and improving integrated care pathways," *ICDM'10 Proceedings of the 10th industrial conference on Advances in data mining: applications and theoretical aspects*, pp. 505–517, 2010.

[129] F. Caron, J. Vanthienen, and B. Baesens, "Healthcare Analytics: Examining the Diagnosis–treatment Cycle," *Procedia Technology*, vol. 9, pp. 996–1004, 2013.

[130] F. Caron, J. Vanthienen, K. Vanhaecht, E. Van Limbergen, J. De Weerdt, and B. Baesens, "Monitoring care processes in the gynecologic oncology department,"

Computers in Biology and Medicine, vol. 44, pp. 88–96, 2014.

[131] F. Caron, J. Vanthienen, J. De Weerdt, B. Baesens, J. De Weerdt, and B. Baesens, “Beyond X-Raying a Care-Flow: Adopting Different Focuses on Care-Flow Mining,” in *the First International Business Process Intelligence Challenge (BPIC11)*, 2011, pp. 1–11.

[132] R. P. J. C. Bose and W. M. P. van der Aalst, “Analysis of Patient Treatment Procedures.,” *Business Process Management Workshops*, vol. 99, pp. 165–166, 2011.

[133] N. G. Weiskopf and C. Weng, “Methods and dimensions of electronic health record data quality assessment: enabling reuse for clinical research,” *JAMIA*, vol. 20(1), pp. 144–151, 2013.

[134] P. Homayounfar, “Process mining challenges in hospital information systems,” in *Federated Conference on Computer Science and Information Systems (FedCSIS)*, 2012, pp. 1135–1140.

[135] H. J. Leavitt, *Applied organizational change in industry, structural, technological and humanistic approaches. Handbook of organizations.* 1965.

[136] J. C. Schlimmer and R. H. Granger, “Beyond Incremental Processing: Tracking Concept Drift,” *AAAI*, pp. 502–507, 1986.

[137] J. Gama, P. P. Rodrigues, and G. Castillo, “Learning with Drift Detection,” *Intelligent Data Analysis*, vol. 3171, no. September, pp. 526–535, 2004.

[138] G. J. Ross, N. M. Adams, D. K. Tasoulis, and D. J. Hand, “Exponentially weighted moving average charts for detecting concept drift,” *Pattern Recognition Letters*, vol. 33, no. 2, pp. 191–198, 2012.

[139] A. Bifet and R. Gavaldà, “Learning from Time-Changing Data with Adaptive Windowing,” *Proceedings of the 2007 SIAM International Conference on Data Mining*, pp. 443–448, 2007.

[140] K. Nishida and K. Yamauchi, “Detecting concept drift using statistical testing,” in *Discovery Science*, 2007, pp. 264–269.

[141] R. P. J. C. Bose, W. van der Aalst, I. Zliobaite, and M. Pechenizkiy, “Dealing With Concept Drifts In Process Mining,” *IEEE Transactions on Neural Networks and Learning Systems*, pp. 1–18, 2013.

[142] D. Kifer, S. Ben-david, and J. Gehrke, “Detecting Change in Data Streams,” *the 30th International Conference on Very Large Data Bases Conference*, pp. 180–191, 2004.

[143] G. Widmer and M. Kubat, “Learning in the presence of concept drift and hidden contexts,” *Machine Learning*, vol. 23, no. 1. pp. 69–101, 1996.

[144] B. F. A. Hompes, J. C. A. M. Buijs, W. M. P. Van Der Aalst, P. M. Dixit, and J. Buurman, “Detecting change in processes using comparative trace clustering,” *CEUR Workshop Proceedings*, vol. 1527, pp. 95–108, 2015.

[145] J. Carmona and R. Gavaldà, “Online Techniques for Dealing with Concept Drift in Process Mining,” *International Symposium on Intelligent Data Analysis*, pp. 90–102, 2012.

[146] N. Kleiner, “Delta analysis with workflow logs: Aligning business process prescriptions and their reality,” *Requirements Engineering*, vol. 10, no. 3, pp. 212–222, 2005.

[147] R. Dijkman, M. Dumas, B. Van Dongen, R. Krik, and J. Mendling, “Similarity of business process models: Metrics and evaluation,” *Information Systems*, vol. 36, no. 2, pp. 498–516, 2011.

[148] S. Kriglstein, G. Wallner, and S. Rinderle-Ma, “A visualization approach for difference analysis of process models and instance traffic,” in *Business Process Management*, 2013, pp. 219–226.

[149] N. R. T. P. van Beest, M. Dumas, L. Garcia-Banuelos, and M. La Rosa, “Log delta analysis: Interpretable differencing of business process event logs,” in *International Conference on Business Process Management*, 2016, pp. 386–405.

[150] A. Bolt, M. De Leoni, and W. M. P. van der Aalst, “A visual approach to spot statistically-significant differences in event logs based on process metrics,” in *International Conference on Advanced Information Systems Engineering*, 2016, pp. 151–166.

[151] M. M. Hennink, *Focus group discussions: understanding qualitative research*. New York, USA: Oxford University Press, 2014.

[152] W. Galitz, *The essential guide to user interface design: an introduction to GUI design principles and techniques*. John Wiley & Sons, 2007.

[153] R. Lenz and M. Reichert, “IT support for healthcare processes - premises, challenges, perspectives,” *Data and Knowledge Engineering*, vol. 61, no. 1, pp. 39–58, 2007.

[154] D. J. Sheskin, *Parametric and non parametric statistical procedures: Third edition*. United States of America: Chapman & Hall/CRC, 2003.

[155] N. Kannan and D. Kundu, *Statistical Signal Processing*. Addison-Wesley Publising Company, 2016.

[156] S. Bersimis, S. Psarakis, and J. Panaretos, “Multivariate statistical process control charts: An overview,” *Quality and Reliability Engineering International*, vol. 23, no. 5. pp. 517–543, 2007.

[157] R. Ghawi, “Process Discovery using Inductive Miner and Decomposition,” Beirut, Lebanon, 2016.

[158] A. Adriansyah, B. F. Van Dongen, and W. M. P. Van Der Aalst, “Conformance checking using cost-based fitness analysis,” *Proceedings - IEEE International Enterprise Distributed Object Computing Workshop, EDOC*, pp. 55–64, 2011.

[159] A. Adriansyah, “Replay a Log on Petri Net for Conformance Analysis Plug-in,” vol. 2012, no. April, pp. 1–15, 2012.

[160] R. Hyndman and G. Athanasopoulos, “Time series decomposition,” in *Forecasting Principles and Practice*, 2nd ed., 2018, pp. 157–182.

[161] J. Anhoj, “qicharts2: Quality Improvement Charts for R,” *The Journal of Open Source Software*, 2019.

[162] A. E. W. Johnson, T. J. Pollard, L. Shen, L. H. Lehman, M. Feng, M. Ghassemi, B. Moody, P. Szolovits, L. A. Celi, and R. G. Mark, “Data Descriptor : MIMIC-III , a freely accessible critical care database,” *Scientific Data*, pp. 1–9, 2016.

[163] A. L. Goldberger, L. A. N. Amaral, L. Glass, J. M. Hausdorff, P. C. Ivanov, R. G. Mark, J. E. Mietus, G. B. Moody, C. Peng, and H. E. Stanley, “Physiobank, physiotoolkit, and physionet components of a new research resource for complex physiologic signals,” *Circulation 101*, vol. 101, no. 23, pp. e215–e220, 2000.

[164] MIT Laboratory for Computational Physiology, “MIMIC-III v1.3,” 2015. [Online]. Available: https://mimic.physionet.org/about/releasenotes/. [Accessed: 03-Aug-2016].

[165] A. Newsham, C. Johnston, and G. Hall, “Development of an advanced database for clinical trials integrated with an electronic patient record system,” *Computers in biology and medicine*, vol. 41, no. 8, pp. 575–586, 2011.

[166] P. H. England, “National Cancer Intelligence Network Older people and cancer About Public Health England,” *Implementation Guide*, vol. 6.0, 2015.

[167] A. P. Kurniati, O. Johnson, D. Hogg, and G. Hall, “Data Quality Issues with Using the MIMIC-III Data for Process Mining in Healthcare,” *Abstract 610 in Scott, P.J. et al. Informatics for Health 2017: Advancing both science and practice, Journal of Innovation in Health Informatics*, vol. 24, no. 1, p. 168, 2017.

[168] A. P. Kurniati, E. Rojas, D. Hogg, and O. Johnson, “The assessment of data quality issues for process mining in healthcare using MIMIC-III , a publicly available e-health record database,” *Health Informatics Journal*, no. 2, 2018.

[169] A. P. Kurniati, G. Hall, D. Hogg, and O. Johnson, “Process Mining in Oncology using the MIMIC-III Dataset,” *IOP Journal of Physics: Conference Series 971*, vol.

971, no. 012008, pp. 1–10, 2018.

[170] Chrisendres, “Online ICD9/ICD9CM codes - Neoplasms,” 2008. [Online]. Available: http://icd9cm.chrisendres.com/index.php?action=child&recordid=1059. [Accessed: 16-Mar-2017].

[171] A. P. Kurniati, G. Hall, D. Hogg, and O. Johnson, “Process mining to explore variation in chemotherapy pathways for breast cancer patients,” *British Journal of Cancer Supplement (Abstract 42)*, vol. 119, no. S1, p. 16, 2018.

[172] Office of National Statistics, “Cancer registration statistics, England:2016,” 2016.

[173] CRUK, “Penile cancer statistics | Cancer Research UK,” *Cancer Research UK*, 2019. [Online]. Available: https://www.cancerresearchuk.org/health-professional/cancer-statistics/statistics-by-cancer-type/prostate-cancer#heading-Four. [Accessed: 01-Aug-2019].

[174] V. I. Levenshtein, “Binary codes capable of correcting deletions,” *Soviet physics doklady*, vol. 10, no. 8, pp. 707–710, 1966.

[175] D. Müllner, “Modern hierarchical, agglomerative clustering algorithms,” no. 1973, pp. 1–29, 2011.

[176] L. Kaufman and P. J. Rousseeuw, *FInding groups in data: an introduction to cluster analysis*, Vol. 344. John Wiley & Sons, 2009.

[177] A. P. Kurniati, C. Mcinerney, K. Zucker, G. Hall, D. Hogg, and O. Johnson, “A multi-level approach for identifying process change in cancer pathways,” in *Process Oriented Data Science for Healthcare*, 2019, pp. 1–12.

[178] A. P. Kurniati, E. Rojas, G. Hall, D. Hogg, and O. Johnson, “Process mining to explore variations in the 62-day pathways of endometrial cancer,” *JCO Clinical Cancer Informatics*, vol. (prepared), 2019.

[179] F. Fox, V. Aggarwal, H. Whelton, and O. Johnson, “A Data Quality Framework for Process Mining of Electronic Health Record Data,” in *IEEE International Conference in Healthcare Informatics (ICHI)*, 2018.

[180] Leeds Teaching Hospitals NHS Trust, “The General Practice Tab-GP Connect,” 2018. [Online]. Available: http://www.ppmsupport.leedsth.nhs.uk/Resources/GPConnect.pdf. [Accessed: 05-Aug-2019].

[181] W. Raghupathi and V. Raghupathi, “Big data analytics in healthcare: promise and potential,” *Health Information Science and Systems*, vol. 2, no. 1, p. 3, 2014.

[182] A. J. M. M. Weijters and J. T. S. Ribeiro, “Flexible Heuristics Miner (FHM),” *BETA working paper series*, vol. 982, pp. 1–5, 2010.

[183] W. Van Der Aalst, A. Adriansyah, and B. Van Dongen, “Causal nets: A modeling language tailored towards process discovery,” *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 6901 LNCS, no. 1, pp. 28–42, 2011.

Appendix A
Training summary

Summary of the planned training, courses, conferences and workshops undertaken during this study are as follow.

Domain A. Knowledge intellectual abilities

A1. Coursera and FutureLearn online courses

#	Course	Host	Grade	Hours
1	Interprofessional Health Informatics	University of Minnesota	100%	8
2	Process Mining: Data Science in Action	Eindhoven Univ of Technology	91%	24
3	Java Programming: Solving Problems with Software	Duke University	92.9%	16
4	Programming Foundations with JavaScript, HTML and CSS	Duke University	81.4%	16
5	Intro to Process Mining with ProM	Eindhoven Univ of Technology	96%	16
6	Process Mining in Healthcare	Eindhoven Univ of Technology	94%	16
7	The Data Scientist's Toolbox	John Hopkins University	96%	16
8	R Programming	John Hopkins University	95%	16
9	Getting and Cleaning Data	John Hopkins University	94%	16
10	Exploratory Data Analysis	John Hopkins University	98%	16
11	Reproducible Research	John Hopkins University	94.9%	16
12	Statistical Inference	John Hopkins University	97.5%	16
13	Regression Models	John Hopkins University	96%	16
14	Applied Plotting, Charting and Data Representation in Python	University of Michigan	98.3%	16
15	Intro to Data Science in Python	University of Michigan	93.8%	16
		Total hours		**260**

A2. YCHI/ LIHS courses

#	Code	Course	Hours
1	YCHI5010m	Informatics in Health Care	40
2	YCHI5015m	The Legal, Ethical and Professional Considerations in Healthcare	40
3	YCHI5030m	Process Modelling, Benefits and Change	40
4	YCHI5045m	Statistics for Health Sciences	40
5	YCHI5055m	Health Data Analytics and Visualisation	40
		Total hours	200

A3. Conferences, workshops, and seminars

#	Conference	Year	Hours
1	Process Mining Camp	2016	8
2	Cancer Data and Outcomes Conference 2016: Using data to drive services	2016	16
3	Health Insights – Regional One Day Event	2016	8
4	The Ideas in Practice Conference – Big Data: Turning Data into Value	2016	8
5	The First Leeds Precision Oncology Symposium	2016	8
6	The 6th Int. Conf. on Information Communication and Management	2016	16
7	2016 NCRI Cancer Conference	2016	16

#	Conference	Year	Hours
8	LIDA Seminar: Process Analytics in Healthcare	2016	4
9	UK Health Data Analytics Network Workshop	2017	16
10	Informatics for Health 2017	2017	24
11	LIDA Seminar: Big Data and Ethics and Pre-seminar Ethics Workshop	2017	2
12	Tableau Workshop	2017	8
13	The WUN Data Science Thematic Workshop	2017	16
14	Cancer Data and Outcomes Conference 2017: Using data to drive services	2017	16
15	Data in Applied Health Research Seminar	2017	2
16	International Conference on Data and Information Science	2017	16
17	LIDA seminar: Introduction to big data in public health	2018	3
18	PHE NCRAS Seminar: Process mining of cancer pathways using EHRs	2018	2
18	The 2018 NCRI Cancer Conference	2018	18
19	UK launch of the International Society for Digital Health	2019	2
20	PODS4H – Workshop Process-Oriented Data Science for Healthcare	2019	8
21	The 2019 Business Process Management Conference	2019	24
	Total hours		**239**

Domain B. Personal effectiveness

#	Course	Host	Hours
1	A balancing act – dealing with the stress of doing a research degree	SDDU	3
2	Time management during your research degree	SDDU	3
3	Preparing for your transfer engineering	SDDU	3
4	Project managing your research degree	SDDU	3
5	Health and Safety Training for DSE Users	UoL	4
6	Fire Safety Training	UoL	4
7	Manual Handling Training	UoL	4
	Total hours		24

Domain C. Research governance and organisation

#	Course	Host	Hours
1	Ethics & ethical review	SDDU	2
2	Ownership, confidentiality and secrecy in research	SDDU	2
3	Research Conduct on VLE	UoL	4
4	CIEH Health & Safety course	UoL	4
5	NIH web-based training on “Protecting Human Research Participants”	MIMIC	8
6	The finishing thesis writer	OD&PL	2
	Total hours		22

Domain D. Engagement influence and impact

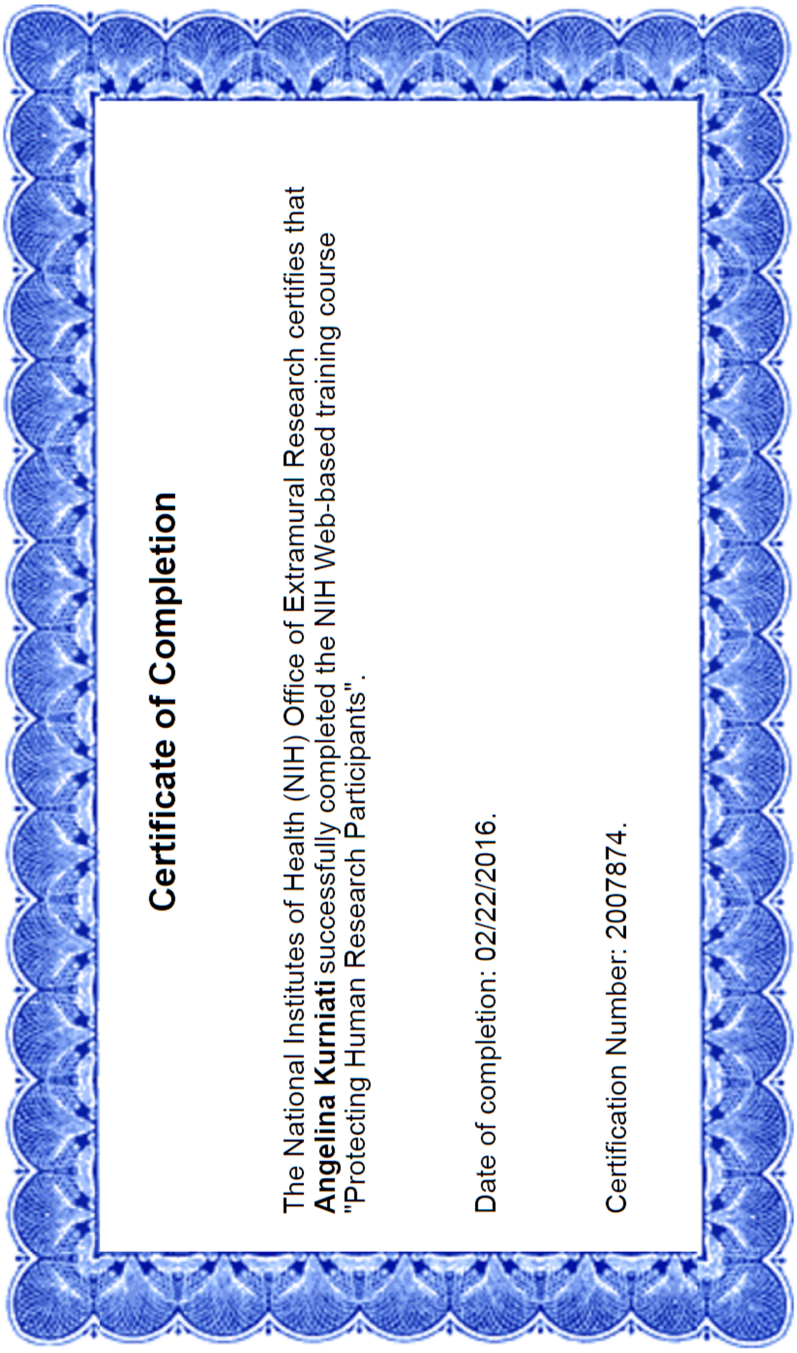
#	Course	Host	Hours
1	Introduction to research impact	SDDU	3
2	Working effectively with your supervisor	SDDU	3
3	Reading critically to write critically PGR workshop	Language	3
4	Foundational Level: Writing purposely workshop	Language	3
5	Giving effective seminar and conference presentations (Science, Engineering and Maths)	SDDU	3
6	Giving effective poster presentations (online)	SDDU	3
7	An Introduction to Effective Research Writing	SDDU	2
	Total hours		20

Appendix B
Ethical approvals

Following are the ethical approvals to access the data used in this study.

B.1 The MIMIC-III Database Access

To access the MIMIC-III database, it is required to complete a web-based training course from the National Institutes of Health (NIH) entitled "Protecting Human Research Participants". The certificate for completing this course is presented below.



Certificate of Completion

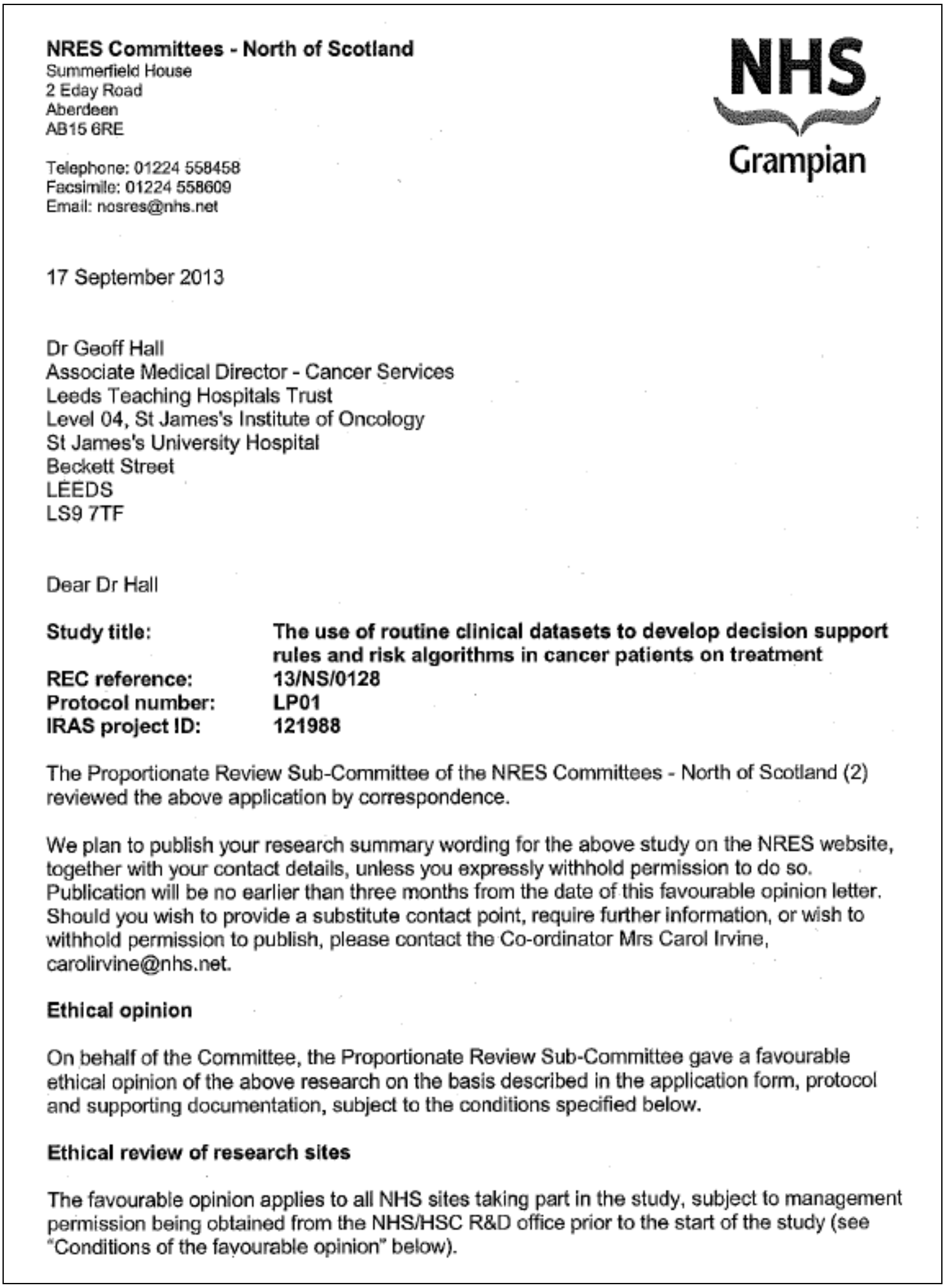
The National Institutes of Health (NIH) Office of Extramural Research certifies that **Angelina Kurniati** successfully completed the NIH Web-based training course "Protecting Human Research Participants".

Date of completion: 02/22/2016.

Certification Number: 2007874.

B.2 The PPM Chemotherapy Extract Access

To access the PPM Chemotherapy extract of the PPM database, I was included in the research team working on this dataset. The study title was "The use of routine clinical datasets to develop decision support rules and risk algorithms in cancer patients on treatment". The first page of the ethical approval for this project is presented below.



NRES Committees - North of Scotland
Summerfield House
2 Eday Road
Aberdeen
AB15 6RE

NHS
Grampian

Telephone: 01224 558458
Facsimile: 01224 558609
Email: nosres@nhs.net

17 September 2013

Dr Geoff Hall
Associate Medical Director - Cancer Services
Leeds Teaching Hospitals Trust
Level 04, St James's Institute of Oncology
St James's University Hospital
Beckett Street
LEEDS
LS9 7TF

Dear Dr Hall

Study title:	**The use of routine clinical datasets to develop decision support rules and risk algorithms in cancer patients on treatment**
REC reference:	**13/NS/0128**
Protocol number:	**LP01**
IRAS project ID:	**121988**

The Proportionate Review Sub-Committee of the NRES Committees - North of Scotland (2) reviewed the above application by correspondence.

We plan to publish your research summary wording for the above study on the NRES website, together with your contact details, unless you expressly withhold permission to do so. Publication will be no earlier than three months from the date of this favourable opinion letter. Should you wish to provide a substitute contact point, require further information, or wish to withhold permission to publish, please contact the Co-ordinator Mrs Carol Irvine, carolirvine@nhs.net.

Ethical opinion

On behalf of the Committee, the Proportionate Review Sub-Committee gave a favourable ethical opinion of the above research on the basis described in the application form, protocol and supporting documentation, subject to the conditions specified below.

Ethical review of research sites

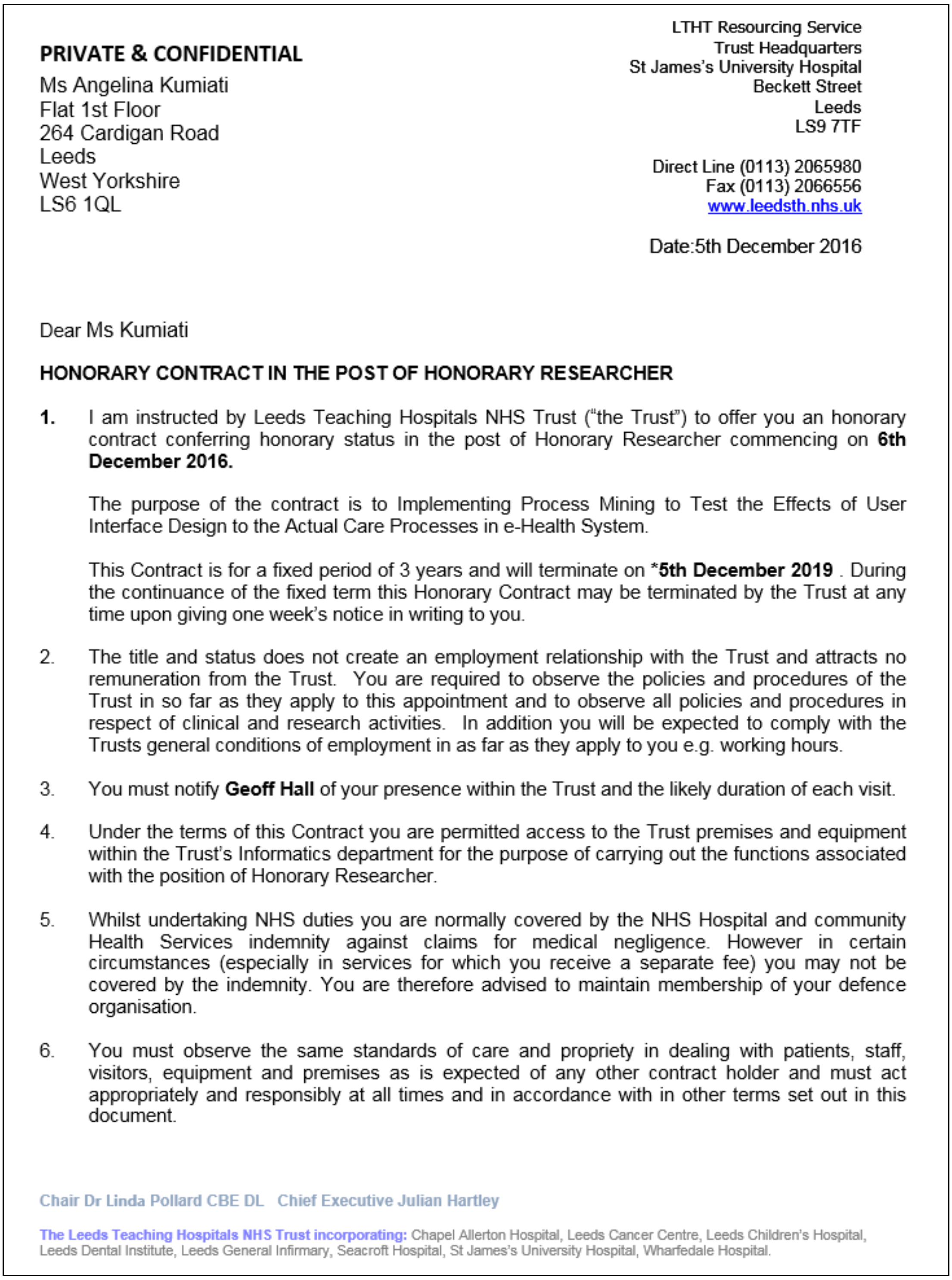
The favourable opinion applies to all NHS sites taking part in the study, subject to management permission being obtained from the NHS/HSC R&D office prior to the start of the study (see "Conditions of the favourable opinion" below).

B.3 The PPM Cancer dataset access

To access the PPM Cancer Dataset, I went through two ethical approvals:

1) **Get NHS LTHT Honorary Contract**

An Honorary Contract is a clinical interaction or period of education or observation which involves Trust employees or patients. The first page of the Honorary Contract for this study is presented below.



LTHT Resourcing Service
Trust Headquarters
St James's University Hospital
Beckett Street
Leeds
LS9 7TF

Direct Line (0113) 2065980
Fax (0113) 2066556
www.leedsth.nhs.uk

Date:5th December 2016

PRIVATE & CONFIDENTIAL

Ms Angelina Kumiati
Flat 1st Floor
264 Cardigan Road
Leeds
West Yorkshire
LS6 1QL

Dear Ms Kumiati

HONORARY CONTRACT IN THE POST OF HONORARY RESEARCHER

1. I am instructed by Leeds Teaching Hospitals NHS Trust ("the Trust") to offer you an honorary contract conferring honorary status in the post of Honorary Researcher commencing on **6th December 2016.**

The purpose of the contract is to Implementing Process Mining to Test the Effects of User Interface Design to the Actual Care Processes in e-Health System.

This Contract is for a fixed period of 3 years and will terminate on ***5th December 2019** . During the continuance of the fixed term this Honorary Contract may be terminated by the Trust at any time upon giving one week's notice in writing to you.

2. The title and status does not create an employment relationship with the Trust and attracts no remuneration from the Trust. You are required to observe the policies and procedures of the Trust in so far as they apply to this appointment and to observe all policies and procedures in respect of clinical and research activities. In addition you will be expected to comply with the Trusts general conditions of employment in as far as they apply to you e.g. working hours.

3. You must notify **Geoff Hall** of your presence within the Trust and the likely duration of each visit.

4. Under the terms of this Contract you are permitted access to the Trust premises and equipment within the Trust's Informatics department for the purpose of carrying out the functions associated with the position of Honorary Researcher.

5. Whilst undertaking NHS duties you are normally covered by the NHS Hospital and community Health Services indemnity against claims for medical negligence. However in certain circumstances (especially in services for which you receive a separate fee) you may not be covered by the indemnity. You are therefore advised to maintain membership of your defence organisation.

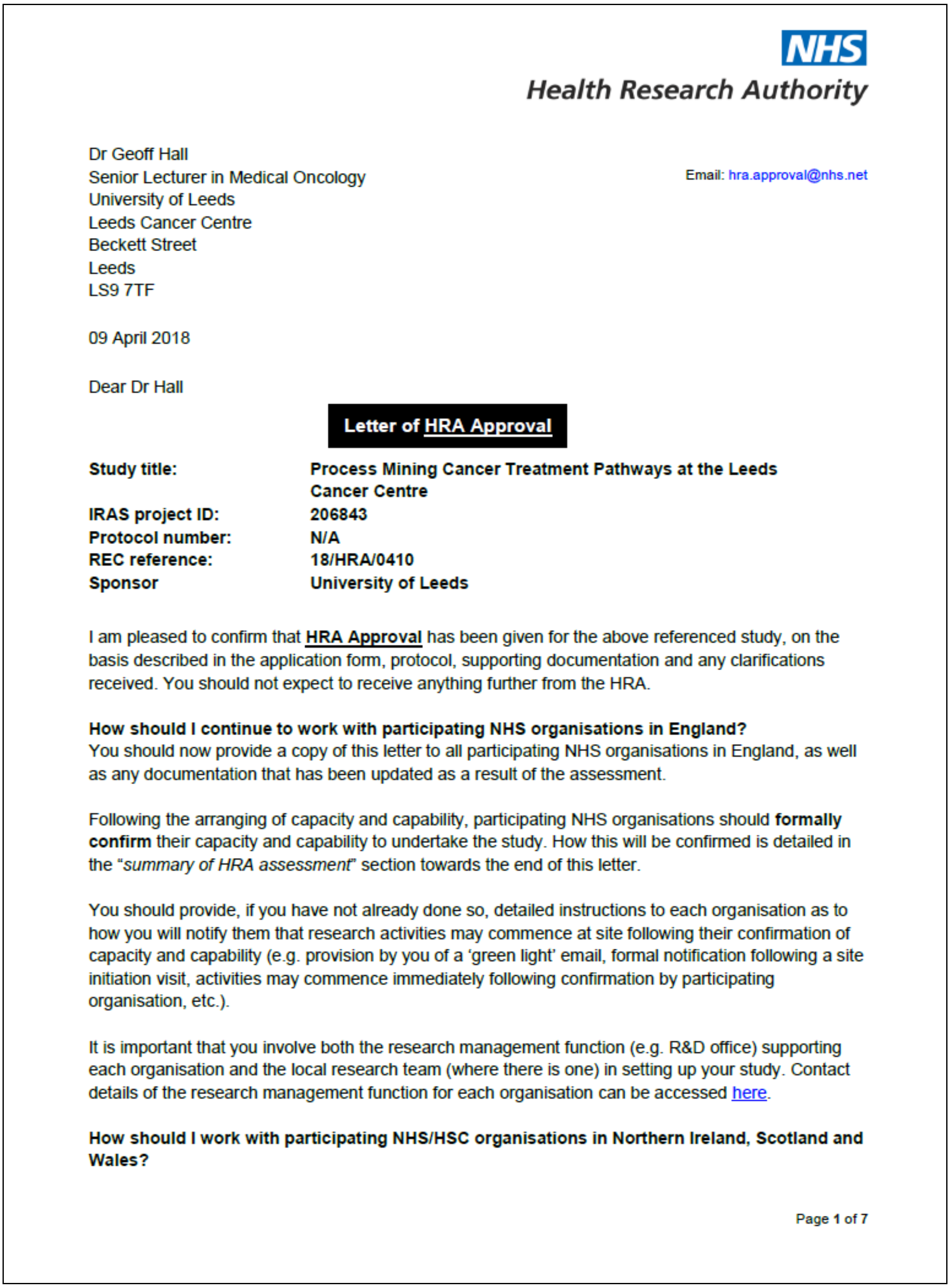
6. You must observe the same standards of care and propriety in dealing with patients, staff, visitors, equipment and premises as is expected of any other contract holder and must act appropriately and responsibly at all times and in accordance with in other terms set out in this document.

Chair Dr Linda Pollard CBE DL   Chief Executive Julian Hartley

The Leeds Teaching Hospitals NHS Trust incorporating: Chapel Allerton Hospital, Leeds Cancer Centre, Leeds Children's Hospital, Leeds Dental Institute, Leeds General Infirmary, Seacroft Hospital, St James's University Hospital, Wharfedale Hospital.

2) IRAS application

The Integrated Research Application System (IRAS) is a single system for applying for the permissions and approvals for health, social and community care research in the UK. This study falls under the requirement to gain the Health Research Authority (HRA) Approval. The first page of the HRA approval is presented below.



NHS
Health Research Authority

Dr Geoff Hall
Senior Lecturer in Medical Oncology
University of Leeds
Leeds Cancer Centre
Beckett Street
Leeds
LS9 7TF

Email: hra.approval@nhs.net

09 April 2018

Dear Dr Hall

Letter of HRA Approval

Study title:	**Process Mining Cancer Treatment Pathways at the Leeds Cancer Centre**
IRAS project ID:	**206843**
Protocol number:	**N/A**
REC reference:	**18/HRA/0410**
Sponsor	**University of Leeds**

I am pleased to confirm that **HRA Approval** has been given for the above referenced study, on the basis described in the application form, protocol, supporting documentation and any clarifications received. You should not expect to receive anything further from the HRA.

How should I continue to work with participating NHS organisations in England?
You should now provide a copy of this letter to all participating NHS organisations in England, as well as any documentation that has been updated as a result of the assessment.

Following the arranging of capacity and capability, participating NHS organisations should **formally confirm** their capacity and capability to undertake the study. How this will be confirmed is detailed in the "*summary of HRA assessment*" section towards the end of this letter.

You should provide, if you have not already done so, detailed instructions to each organisation as to how you will notify them that research activities may commence at site following their confirmation of capacity and capability (e.g. provision by you of a 'green light' email, formal notification following a site initiation visit, activities may commence immediately following confirmation by participating organisation, etc.).

It is important that you involve both the research management function (e.g. R&D office) supporting each organisation and the local research team (where there is one) in setting up your study. Contact details of the research management function for each organisation can be accessed here.

How should I work with participating NHS/HSC organisations in Northern Ireland, Scotland and Wales?

Page **1** of **7**

Appendix C
Experiments of the MIMIC-III dataset

C.1 Data quality assessment of the MIMIC-III dataset

UNIVERSITY OF LEEDS School of Computing	**EXPERIMENT DOCUMENTATION**	**Date of experiment** 01/05/2017
	Experiment title: Data quality assessment of process mining using MIMIC-III database	**Experiment code** M3-DQ-ALL001
	Researcher's name: Angelina Kurniati	

Area of investigation

This experiment is a data quality assessment of MIMIC-III dataset for process mining. The complete report of this experiment has been published in a paper in the Health Informatics Journal.

Data source

The MIMIC-III dataset is analysed per table.

Research questions

(1) Can the MIMIC-III database be used to better understand data quality issues for process mining in healthcare?

(2) What are the data quality issues for process mining with MIMIC-III?

(3) How might the change in the EHR system in 2008 affect the data quality?

Hypothesis

Weiskopf &Weng (W&W) framework can be used to identify data quality issues for process mining with MIMIC-III.

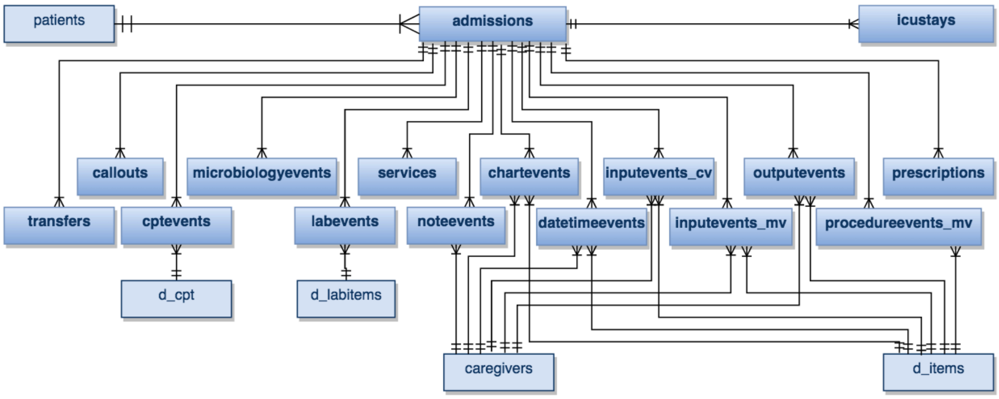
Method

The general method used in this experiment is an adaptation of the L* lifecycle model, which includes data quality assessment. The methodology consists of *Plan and justify (Stage 0), Database reconstruction (new stage), Extract (Stage 1), Create a control-flow model (Stage 2), Create integrated process model (Stage 3), and Data quality assessment (new stage)*. Database reconstruction was added to support Data quality assessment and trigger the next iterations.

Result and Discussion

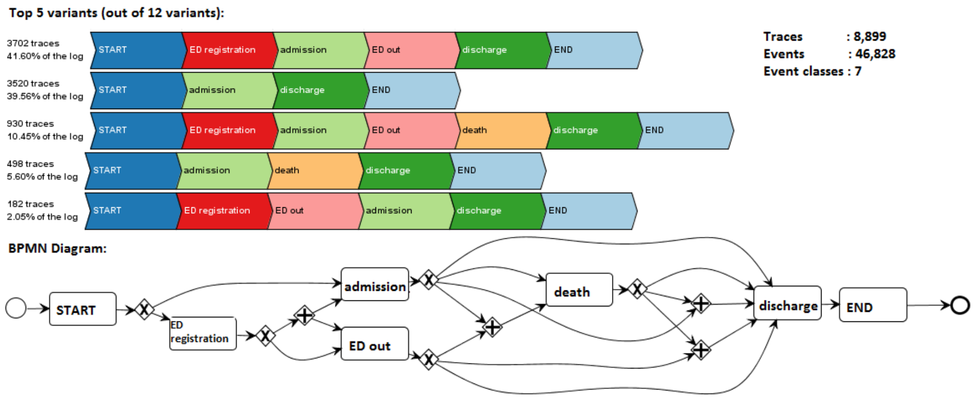
1) **Planning and justification** have been done by specifying three research questions. Q1 was addressed through a cancer-specific example, Q2 was addressed by applying the Weiskopf& Weng framework, and Q3 was addressed by investigating the differences of processes in CV and MV systems.

2) **Database reconstruction** was needed as a foundation to support iterative experiments on data quality assessment. The downloaded data were in .csv files (26 files, more than 6 GB in total) and were imported to a PostgreSQL database. The concept level Entity-Relationship Diagram (ERD) is presented in the following Figure.



The entities in bold contain timestamped information which can be used to construct an event log for process mining.

3) **Extraction, transformation, and loading (ETL).** Extraction was done multiple times from the reconstructed database. For example, the admission table in the MIMIC-III database was extracted by selecting [`admittime, dischtime, edregtime, edouttime, deathtime`] of cancer patient admissions [icd9_codes 140x-239x]. Process discovery was done using three plugins in ProM: (1) Convert CSV into XES, (2) Add artificial events >> START and END events, and (3) BPMN Analysis using Heuristics Miner. The Q1 (*Can the MIMIC-III database be used to better understand data quality issues for process mining in healthcare?*) was addressed to find the most followed admission paths of cancer patients. The five most common variants and BPMN process model of admissions are presented in the following Figure.



4) **Process analytics** was done by extending with time and resource perspective. In this experiment, it was done to study the effect of EHR change, which addressed Q3 (*How does the change in the EHR system in 2008 affect the data quality?*). The approach was to work backwards from the *inputevents_cv* and *inputevents_mv* tables to identify which hospital admissions has been recorded on which EHR. We also identified the differences through four tables: *chartevents, datetimeevents, inputevents,* and *outputevents*. Each admission was marked with CV, MV, or both; then we created separate event logs from each data sources. We then compared the models using DifferenceGraph plugin in ProM.

5) **Data quality assessment** was done based on the W&W framework, which has been summarised in Table 4.1 in Section 4.1.3.

Conclusion

Despite some limitations found in this data quality assessment, the MIMIC-III database was found to be sufficient for process mining projects. The three minimum components required for process mining are available in 16 tables in the MIMIC-III database. Process miners can also specify different levels of details needed in the analysis.

C.2 Log profiling for CV-MV comparison

UNIVERSITY OF LEEDS School of Computing	**EXPERIMENT DOCUMENTATION**	**Date of experiment** 07/03/2018
	Experiment title: Data profiling of MIMIC-III dataset for CV-MV comparison	**Experiment code** M3-CVMV-PRF001
	Researcher's name: Angelina Kurniati	

Area of investigation

This experiment is data profiling of MIMIC-III dataset as an initial step for CV-MV comparison.

Data source

The MIMIC-III dataset is analysed per table.

Research question

Is analysing each table in the MIMIC-III dataset provides initial information for CV-MV comparison?

Hypothesis

Data profiling would give initial information for CV-MV comparison.

Method

1) **Extraction** has been done by database reconstruction process in PostgreSQL.

2) **Create** event logs of each event tables with [case_id, activity, timestamp, tsource] format, with:

```
case_id = subject_id/ hadm_id/ icustay_id
activity = {identified in Table 4.5}
timestamp = {column named -time in each table}
```

2) Save as .csv files

3) **Load** into DISCO and ProM (if possible), and analyse the result.

Result and Discussion

Data profiling was done to analyse the similarities and differences of patient data in CV and MV systems. This analysis was done per table in the MIMIC-III database. Only data profiles that potentially related to process mining are presented here.

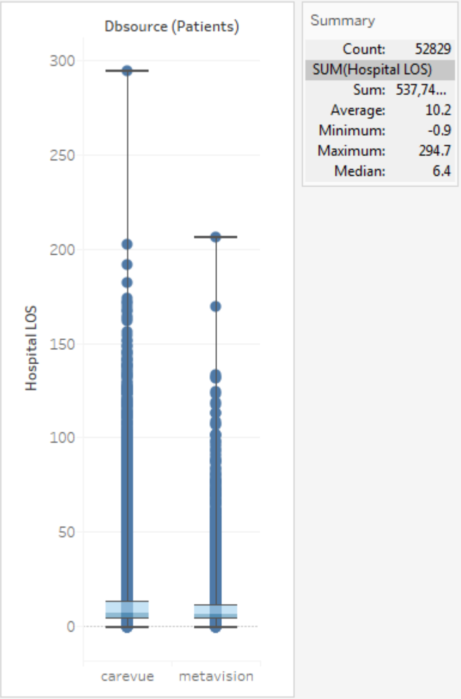
1. Admissions

This table sourced from hospital databases, to define a patient’s hospital admission, HADM_ID.

- NEWBORN admissions only happened in CV (7,863 admissions/ 23.46%). It has also been described in the data descriptor that NEWBORN admissions were not included in the MV extracts. It was due to the limitation of the de-identification procedure to maintain confidentiality.

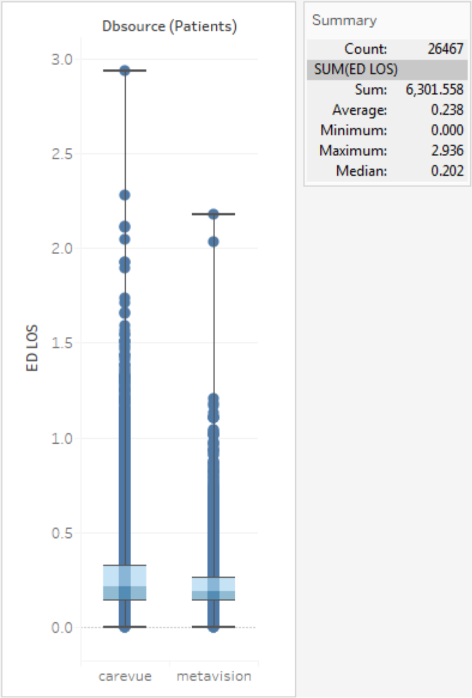
EHR System	Admission type	Admissions	%
CV	Elective	3,983	11.88
	Emergency	20,649	61.61
	Newborn	7,683	23.46
	Urgent	1,020	3.04
MV	Elective	2,960	15.33
	Emergency	16,130	83.51
	Urgent	224	1.16

- There were data quality issues detected in CV where ED duration < 0 (98 admissions). After non-valid data excluded, the average ED durations in MV (0.212 days) are shorter than CV (0.238 days). It means that ED duration has been shortened in the new system.



Summary

Count:	52829
SUM(Hospital LOS)	
Sum:	537,74...
Average:	10.2
Minimum:	-0.9
Maximum:	294.7
Median:	6.4



Summary

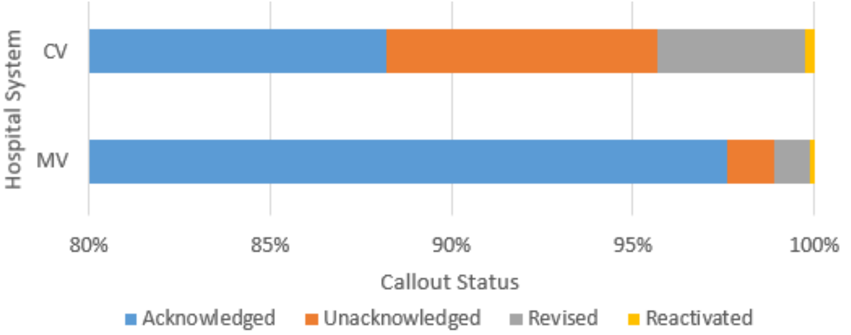
Count:	26467
SUM(ED LOS)	
Sum:	6,301.558
Average:	0.238
Minimum:	0.000
Maximum:	2.936
Median:	0.202

- There are 204 records without admission location in CV, but there is none in MV. It means that data quality is improved in the new system, in term of recording admission location.

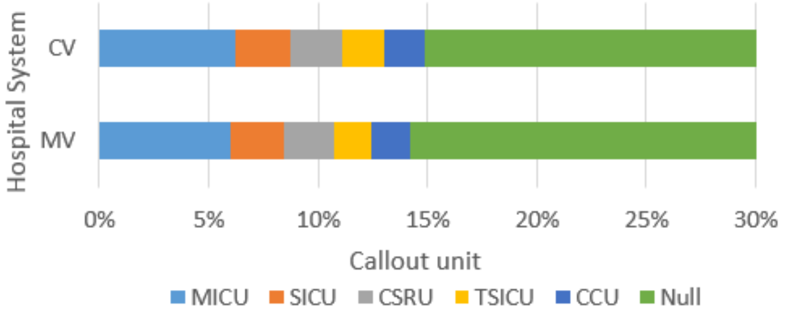
2. callout

Provides information when a patient was READY for discharge from the ICU, and when the patient was actually discharged from the ICU.

- Acknowledge status are mostly *Acknowledged* both in CV (8,078/ 91.61%) and in MV (15,560/ 99.01%).

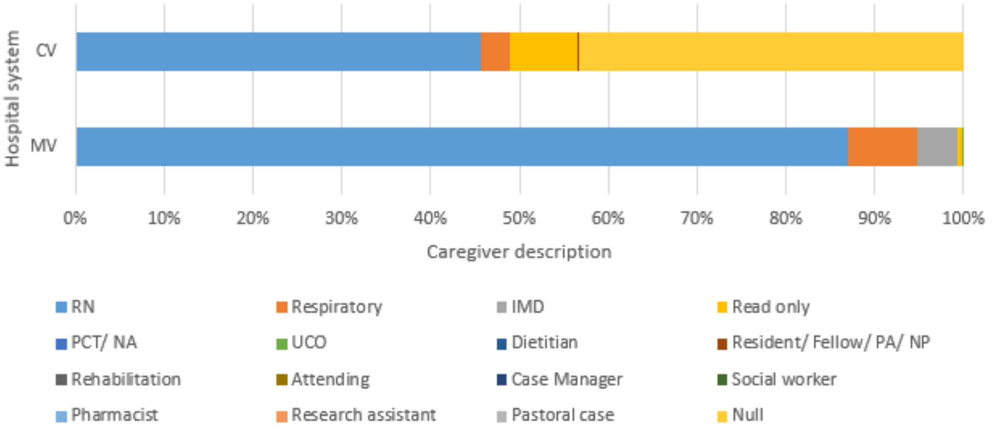


- Both in CV and MV, the three most common callout units are MICU, SICU, and CSRU. In both systems, callout units were rarely recorded, with *Null* value of 8,208 (93.08%) in CV and 14,914 (94.90%) in MV.



3. caregivers

Defines the roles of caregivers in CV and MV databases. Some caregiver types are only available in MV, which are: attending, pastoral care, rehabilitation, research assistant, and social worker. There are a significant number of NULL values in CV (43.25%), but none in MV.



4. chartevents

Contains all charted data for all patients in CV and MV databases.

Characteristic	CV	MV
Distinct subject_id	25,505	16,116
Distinct hadm_id	29,541	19,248
Distinct icustay_id	31,329	20,582
Distinct itemid	5,100	1,868
Distinct cgid	1,177	1,358

5. cptevents

Contains current procedural terminology (CPT) codes, which facilitate billing for procedures performed on patients.

Characteristic	CV	MV
Distinct subject_id	16,279	15,856
Distinct hadm_id	19,588	19,250
Distinct costcenter	"ICU"167737 "Resp";57934	"ICU";242727 "Resp";34406
Distinct cpt_cd	971	1786
Distinct cpt_number	970	1780
Distinct cpt_suffix	0	1
Distinct ticket_id_seq	550	531
Distinct description	4	4

6. d_items

This is the reference table to separate CV and MV records in the other tables. Information available in the d_items table are:

Characteristic	CV	MV
Distinct itemid	9059	2992
Distinct abbreviation	0 (all NULL)	2907
Linksto	chartevents, datetimeevents, inputevents_cv, outputevents	chartevents, datetimeevents, inputevents_mv, outputevents, procedureevents_mv
Distinct category	23	68
Distinct unitname	0 (all NULL)	53

Parameter types in MV system: checkbox (c), datetime (d), numeric (n), numeric with tag (nt), process (p), solution (s), and text (t).

Linksto	c	d	n	nt	p	s	t
chartevents	v		v	v			v
datetimeevents		v					
inputevents_mv						v	
outputevents		v	v				v
procedureevents_mv					v		

7. d_labitems

This is the reference table to separate CV and MV records in the labevents table. Information available in the d_labitems table are:

Characteristic	CV	MV
Distinct itemid	705	675
Distinct label	557	534
Distinct fluid	13	12
Distinct category	6	6
Distinct loinc_code	572	552

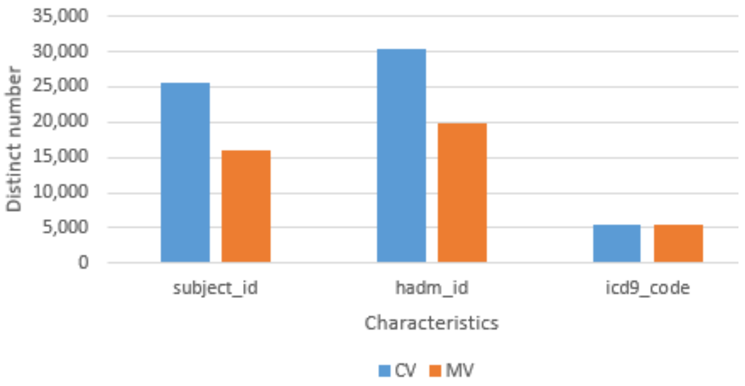
8. datetimeevents

Contains all date formatted data in the CV and MV databases. Column 'resultstatus' is Null in all rows. Columns 'warning' and 'error' are mostly Null in CV, but not in MV.

Characteristic	CV	MV
Distinct subject_id	11395	16116
Distinct hadm_id	12759	19242
Distinct icustay_id	13495	20559
Distinct itemid	28	131

9. diagnosis_icd

This table contains ICD diagnoses for patients in ICD-9 codes, which were generated for billing purposes at the end of the hospital stay.



10. **drgcodes**: contains diagnosis-related groups (DRG) codes for patients. *per admission

DRGs are used to categorise inpatient hospital visits severity of illness, risk of mortality, prognosis, treatment difficulty, need for intervention, and resource intensity. The DRG system was developed for statistical classification of hospital cases. (*icd.codes/articles/what-is-drg*)

Characteristic	CV	MV
Distinct drg_code	1515	1512
Distinct description	1126	940

* *icd.codes/articles/what-is-drg*

11. icustays

This table defines each ICUSTAY_ID in the database. The ICUSTAYS table is derived from the TRANSFERS table, grouped based on ICUSTAY_ID, and excluded rows without ICUSTAY_ID.

Characteristic	CV				MV			
	Unit	#sid	#hid	#iid	Unit	#sid	#hid	#iid
First_careunit	CCU	4024	4328	4455	CCU	2165	2320	2373
	CSRU	5040	5260	5478	CSRU	2909	3053	3189
	MICU	7555	9025	9452	MICU	6793	8121	8521
	NICU	4795	4914	4984	SICU	3504	3775	3975
	SICU	3610	3875	4082	TSICU	2527	2599	2702
	TSICU	3133	3205	3298				
Last_careunit	Unit	#sid	#hid	#iid				
	CCU	4024	4328	4455				
	CSRU	5040	5260	5478				
	MICU	7555	9025	9452				
	NICU	4795	4914	4984				
	SICU	3610	3875	4082				
	TSICU	3133	3205	3298				

12. inputevents

Inputevents in the CV system is in inputevents_cv table and inputevents in the MV system is in inputevents_mv table. Observations in those two tables are not duplicated and can be unioned to create complete inputevents when needed. Only the CHARTTIME is available in CV, while STARTTIME and ENDTIME are available in MV. The CHARTTIME in CV is correspond with starttime.

Characteristic	CV	MV
Distinct subject_id	25,450	16,012
Distinct hadm_id	29,353	19,111
Distinct icustay_id	31,011	20,393
Distinct itemid	2,794	274
Distinct cgid	1,078	500

13. labevents

This table contains all laboratory measurements for a given patient, including outpatient data. Lab measurements for outpatients do not have a HADM_ID. The item identifiers can be explained with reference to *d_labitems* table.

Characteristic	CV	MV
Distinct itemid	705	675
Distinct subject_id	25404	16087
Distinct hadm_id	30040	19581

Additional columns are *value*, *valuenum*, *valueuom*, and *flag*. *Value* contains the value measured for the itemid. If the *value* is numeric, *valuenum* contains the same data in numeric format, otherwise *valuenum* is null. *Valueuom* is the unit measurement for the value. *Flag* indicates whether the value is abnormal or not.

14. microbiologyevents

This table contains microbiology information, including tests performed and sensitivities.

Characteristic	CV	MV
Distinct subject_id	19,454	16,420
Distinct hadm_id	22,840	18,574
Distinct spec_itemid	89	81
Distinct org_itemid	282	215
Distinct spec_itemid	30	29

15. noteevents

This table contains all notes for patients. There is no HADM_ID for outpatients and for inpatients who was not admitted to the ICU for that particular hospital admission. The main part of this table is the text, which contains all notes for patients and is potential to be analysed with text analytics.

Characteristic	CV	MV
Distinct subject_id	25,532	15,867
Distinct hadm_id	30,376	19,445
Distinct category	13	15
Distinct description	2,025	2,812
Distinct cgid	1,215	1,087
Distinct text	1,175,306	634,789

16. outputevents

This table contains output data for patients. Metavision ITEMID values are all above 220000, while a subset of commonly used medications in CareVue data have ITEMID 30000-39999. ISERROR is a Metavision checkbox where a caregiver can specify that an observation is an error.

Characteristic	CV	MV
Distinct subject_id	24,251	15,932
Distinct hadm_id	27,951	18,951
Distinct icustay_id	29,405	20,200
Distinct itemid	1,035	70
Distinct cgid	1,038	837

17. patients

This table contains all charted data for all patients. DOB has been shifted for patients older than 89. The median age for the patients whose date of birth was shifted is 91.4. DOD is the date of death for the given patient, merged from DOD_HOSP (from the hospital database) and DOD_SSN (from the social security database), giving priority to DOD_HOSP if both were recorded. Patients table contains 25,534 CV patients and 16,116 MV patients.

Gender	CV	MV
M	14,491	9,040
F	11,043	7,075

Expire flag	CV	MV
0	15,358	11,397
1	10,176	4,718

18. prescriptions

This table contains medication related to prescriptions, sourced from the hospital provider order entry database. DRUG_TYPE provides the type of drug prescribed. DRUG, DRUG_NAME_POE, DRUG_NAME_GENERIC are various representations of the drug prescribed to the patient. FORMULARY_DRUG_CD, GSN, NDC provide a representation of the drug in different coding systems. GSN is the Generic Sequence Number, NDC is the National Drug Code. ROUTE is the route prescribed for the drug.

Characteristic	CV	MV
Distinct subject_id	20,801	16,096
Distinct hadm_id	24,574	19,587
Distinct icustay_id	25,393	20,544
Distinct drug	3,253	2,700
Distinct prod_strength	3,245	2,139
Distinct route	67	71

19. procedureevents

Contains procedures for patients in MV database.

Characteristic	Distinct values
subject_id	16,024
hadm_id	19,124
icustay_id	20,405
itemid	116
location	96
cgid	957

20. procedures_icd

This table contains ICD procedures for patients in ICD-9 procedure codes. The ICD codes are generated for billing purposes at the end of the hospital stay. Seq_num provides the order of the performed procedures. ICD9_code can be joined to the d_icd_procedures table to get the descriptions.

Characteristic	CV	MV
Distinct subject_id	23,601	14,122
Distinct hadm_id	27,832	16,898
Seq_num	1-40	1-40
Distinct icd9_code	1,719	1,557

21. services

This table lists services under which a patient was admitted/ transferred under. While a patient can be physically located at a given ICU type, they are not necessarily being cared for by the team which staffs that ICU.

Characteristic	CV	MV
Distinct subject_id	25,520	16,110
Distinct hadm_id	30,484	19,736
Distinct prev_service	17	17
Distinct curr_service	20 *Not in MV: NB (Newborn at hospital), TRAUM (trauma), NBB (newborn baby), PSYCH (psychiatric)	17 *Not in CV: -

22. **transfers**

This table records physical locations for patients throughout their hospital stay. The ICUSTAYS table is derived from this table. ICUs have moved throughout the years, same WARDID may be an ICU for patient A but not an ICU for patient B.

EVENTTYPE describes what transfer event occurred, which includes transfer, admit, and discharge. There are 15 events in CV with eventtypes = NULL, but none in MV.

Characteristic	CV	MV
Distinct subject_id	25,534	16,115
Distinct hadm_id	30,514	19,748
Distinct icustay_id	31,749	20,759
Distinct prev_wardid	53	45
Distinct curr_wardid	53	45
LOS	81.465 [0 – 20,880]	63.029 [0 – 7,612.52]

Conclusion

The log-based comparison reveals some important information to consider in the next step in process mining to analyse process changes in CV and MV systems in the MIMIC-III database. Differences in columns and data details recorded in the CV and MV systems should be considered when comparing process in those two systems. For example, newborn admissions were only included in the CV system and not in MV. The consequence is when the comparison is done to all records in the CV and MV systems, they will not be comparable. Calculation of the average age at admission will also include newborn admissions in the CV system and not in the MV system.

Another insight gained from this experiment is that it is not valid to analyse all data in the MIMIC-III database without considering two different systems from which the admissions were recorded. Data profiling of those two EHR systems revealed many differences in the data level that affected the process analysis.

C.3 Model-based comparison of CV-MV

UNIVERSITY OF LEEDS School of Computing	**EXPERIMENT DOCUMENTATION**	**Date of experiment** 18/03/2018
	Experiment title: Model comparison of MIMIC-III for CV-MV	**Experiment code** M3-CVMV-MDL001
	Researcher's name: Angelina Kurniati	

Area of investigation

This experiment is model comparison of MIMIC-III tables as a part of CV-MV comparison.

Data source

The MIMIC-III dataset is analysed per table.

Research question

Is comparing process models of each table in the MIMIC-III dataset provides initial information for CV-MV comparison?

Hypothesis

Model comparison would give control-flow change information for CV-MV comparison.

Method

1) **Extraction** has been done in database reconstruction process in PostgreSQL.

3) **Create** event logs of each event tables with [case_id, activity, timestamp, tsource] format, with:

case_id = subject_id/ hadm_id/ icustay_id

activity = {identified in M3-TRS-DES001}

timestamp = {identified in M3-TRS-DES001}

2) Save as .csv files

3) **Load** into DISCO and ProM (if possible)

- Merge subsequent events
- Add artificial START and END events
- Create PN with Heuristics Miner

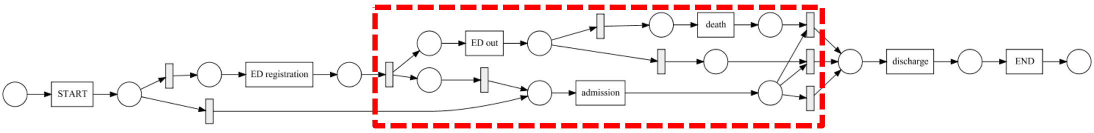
4) Analyse the results

Result and Discussion

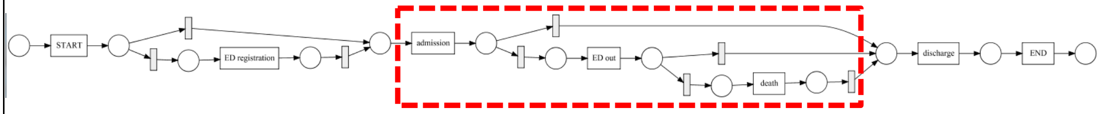
Model comparisons were done by creating a process model for each table in CV and MV. This is expected to reveal control-flow changes in CV and MV systems.

1. **Admissions**

CV



MV

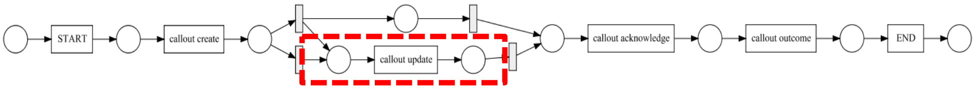


Differences

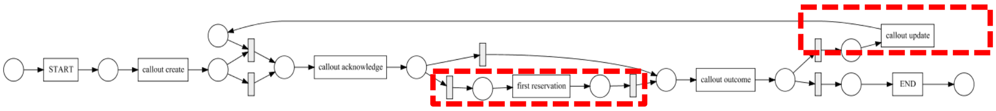
- In CV, ED out can happen before or after admission. In MV, admission happened as an optional activity before ED out.

2. **callout**

CV



MV



Differences

- There are four activities in CV and five activities in MV. The one additional activity in MV is *first reservation*. This additional activity happens after *callout acknowledge* and before *callout outcome*.
- In CV, *callout update* is an optional activity happened after *callout create* and before *callout acknowledge*. In MV, callout update happened in the backward path after *callout outcome* and before *callout acknowledge*.

Conclusion

The model comparison reveals some important information to consider in the next step in process mining to analyse process changes in CV and MV systems in the MIMIC-III database. The limitation of this approach is that it is heavily dependent on the visual differences of the process models. There was an opportunity to find a more structural/ quantitative way to analyse the differences in the properties of two process models.

Appendix D
Experiments of the PPM Chemotherapy dataset

D.1 Data description of the PPM Chemotherapy dataset

UNIVERSITY OF LEEDS	**EXPERIMENT DOCUMENTATION**	**Date of experiment** 13/02/2017
	Experiment title: PPM Chemotherapy Data Description	
School of Computing	**Researcher's name:** Angelina Kurniati	**Experiment code** CHE-ALL000

Area of investigation

This experiment is for understanding PPM Chemotherapy dataset through building data description, which include creating an event log from each table in the database, with the minimum format is [*case_id, activity, timestamp*].

Data source

The dataset is a copy of DB2-2016 database, which is an anonymous dataset of cancer patient treatment in Leeds Cancer Centre during 1996-2015. This database was originally created to develop decision support rules and risk algorithms in cancer patients on treatment.

Research question

Is the PPM Chemotherapy database can be used for process mining?

Hypothesis

The tables in the PPM Chemotherapy database can be used as it provides at least minimum requirements for process mining.

Method

1) **Extract** by copying the DB2-2016 database in SQL Server from all potentially useful tables.
2) **Describe** each table for understanding each attribute and the relations between attributes. This includes identifying columns needed for process mining (especially activities) and identifying splitting attributes for next experiments.
3) **Transform** to create event log with [case_id, activity, timestamp] format. For example, admissions table has been transformed into an event log with:

   case_id = PID/ KTPId

   activity = {*admission, discharge*}

   timestamp = {*GenDateAtAdmission, GenDateAtDischarge*}

* all timestamp attributes are created by generating dates based on ages (in number of days) with ’01-01-2020’ as the date of birth for all patients.

4) Save as .csv files

5) **Load** into DISCO and ProM (if possible), and analyse the result.

Result and Discussion

1) Admissions

The ADMISSIONS table gives information regarding the patient’s admission to the hospital. It is linked to other tables through PID and AdmissionId.

Table columns:

Column	Data type	Distinct	Missing	Description
PID	unique id	28,609	0	Case id
AdmissionId	int	397,444	0	
AgeAtAdmission	int	33,388	0	Activity
ContactSpecialityCode	int	80	910	
ContactSpecialityLabel	nvarchar(100)	80	910	
AdmissionMethod	varchar(9)	3	0	
MethodCode	Int	17	0	
MethodLabel	nvarchar(100)	17	0	
AgeAtDischarge	int	33,365	25	0 – 36,649
Number of rows		397,444		

Notes:

(1) Potential case_id = PID;

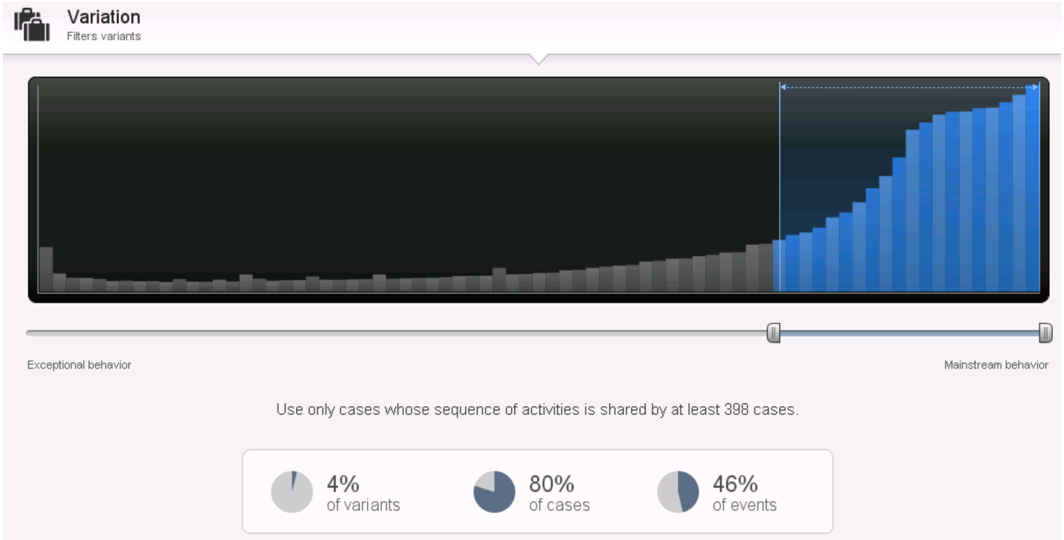
(2) AgeAtDischarge=0 in 25 admissions;

(3) Potential activities: ‘Admission’ and ‘Discharge’ or AdmissionMethod/ MethodLabel

Details:

Cases	: 28,609
Events	: 794,863
Events per case	: min 2, mean 28, max 1020
Mean case duration	: 35.3 months
Median case duration	: 17.1 months
Case duration	: 0 days – 12 years 11 months 23 days
Variants	: 500

Variance checking shows that 80% of cases can be represented by 4% of variants:



Use only cases whose sequence of activities is shared by at least 398 cases.

The admission table can be used for process mining, providing that this table contains the minimum requirements for process mining.

2) *AdmissionWardStayLocation*

The AdmissionWardStayLocation table gives detailed information regarding the patient's admission to the hospital, ward stays location, start and end of staying. It is linked to other tables through PID and AdmissionId.

Column	Data type	Distinct	Missing	Description
PID	unique id	28,609	0	Case id
AdmissionId	int	397,423	0	
AgeAtAdmissionDischarge	int	33,365	37	0-36,649
AgeAtWardStayStart	int	33,730	0	0-36,642
AgeAtWardStayEnd	int	33,748	7	0-36,649
ew_WardLabel	nvarchar(100)	360	0	
Number of rows		469,886		

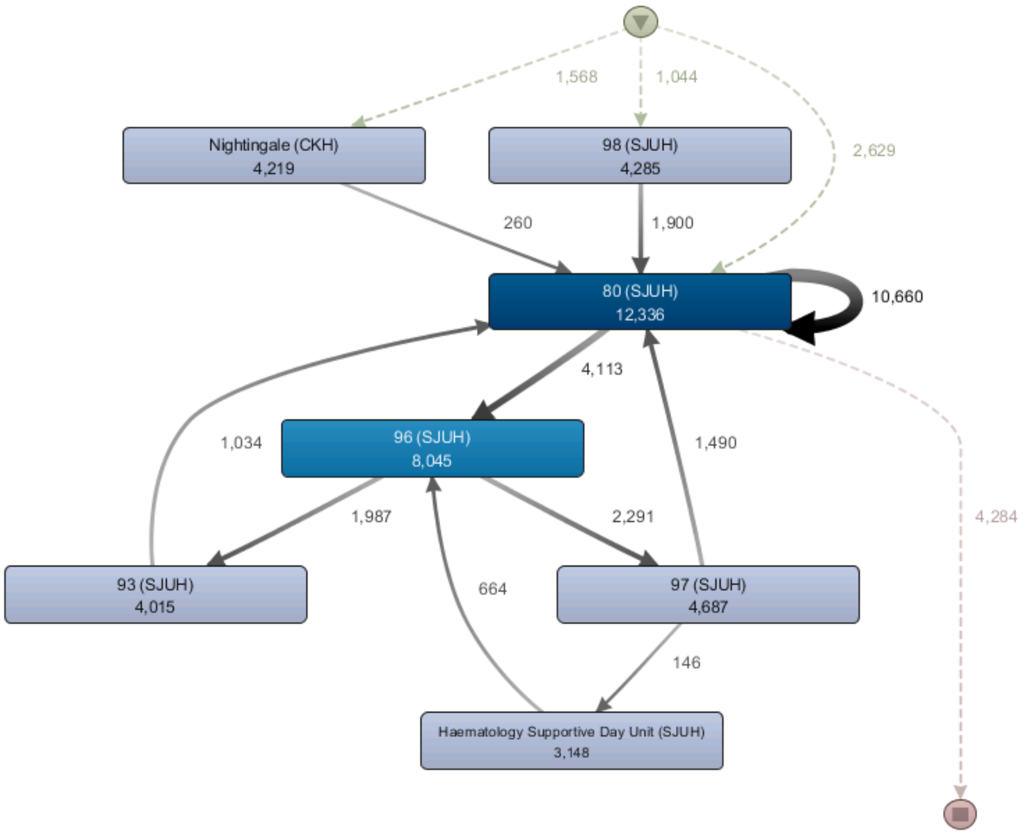
Two important points are:

(1) AgeAtAdmissionDischarge is identical with AgeAtDischarge in the Admissions table.
(2) Potential activitiy is WardLabel.

Details:

Events	: 469,886
Cases	: 28,609
Events per case	: min 1, mean 16.5, max 517
Activities	: 360
Mean case duration	: 35.3 months
Median case duration	: 17.1 months
Case duration	: 0 days - 12 years 11 months 23 days
Variants	: 24,460

Fuzzy model:



(1.7% activities, 0% paths, showing case frequency)

The fuzzy model above shows the most frequently assigned ward is ward 80 (n=12,336), followed by ward 96 (n=8,045) and ward 97 (n=4,687). A further search has been done and found that ward 80 is the oncology ward, ward 96 is an acute assessment ward, and ward 97 is an oncology ward for patients with different types of cancer.

3) ChemoCycles

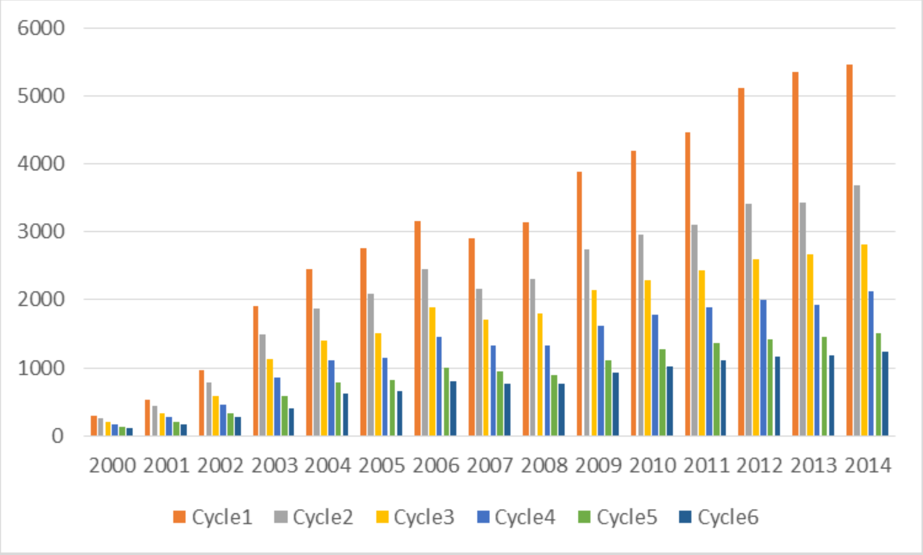
The ChemoCycles table provides detail information regarding chemotherapy cycles of the patient. This is linked to other tables through PID and RegimenID.

Column	Data type	Distinct	Missing	Description
PID	unique id	29,009	0	PID-Cycle = 1-m
CycleID	int	198,096	0	RegimenID-CycleID = 1-m
RegimenID	int	56,624	0	PID-RegimenID = m-m
AgeWhenCycleStarted	int	28,399	0	2-36,626
YearCycleStarted	int	20	0	1996-2015
CycleNumber	int	93	0	1-93
CycleMaxDays	int	131	0	-7 – 547
CycleCalculatedLength	int	131	0	-7 – 547
Number of rows		198,096		

Notes: (1) Potential case_id: PID, CycleID, RegimenID

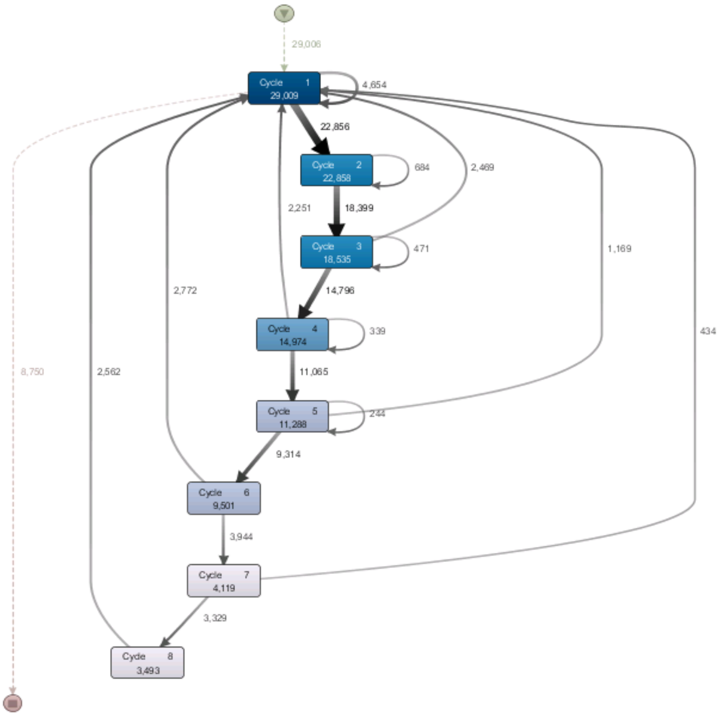
(2) Potential activities: CycleNumber

Number of cycles started per year:



The diagram above shows that the number of cycle started is generally increasing over the years, but the number of subsequent cycles is decreasing. It has been further analysed in the experiment 3 using the PPM Chemotherapy dataset as presented in Section 5.2.

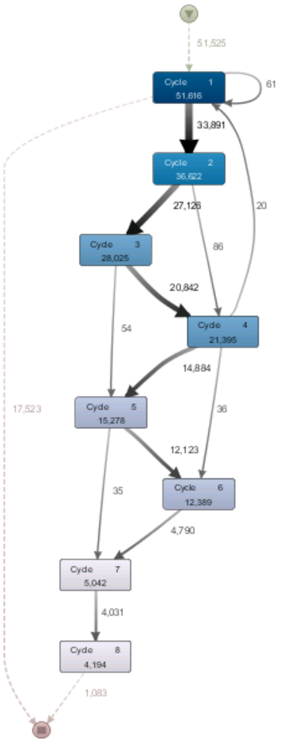
Fuzzy models:



(8% activities, 1.7% paths)

(case frequency)

(caseID = PID)



(8% acts, 2.1% paths)

(case freq)

(caseID = PID+RegimenID)

* The left model shows model with caseID = PID, which is not fair, because a patient might have more than one RegimenID and the ChemoCycles might represent cycles of different regimen. This setting would be needed to combine this event log with event logs from other tables.

* The right model shows model with caseID = PID+RegimenID, which represent ChemoCycles for each regimen of each patient. It shows that patients are generally followed subsequent cycles, with 61 patients (0.12%) repeated cycle 1 and 20 patients (0.04%) had to go back to cycle 1 from cycle 4.

Details:

Events	: 198,096
Cases	: 56,624
Events per case	: min 1, mean 3, max 91
Activities	: 93
Mean case duration	: 78.3 days
Median case duration	: 42 days
Case duration	: 0 days - 6 years 6 months 17 days
Variants	: 1,105

4) ChemoDrugs

The ChemoDrugs table provides detail information regarding chemotherapy drugs given to the patients. This is linked to other tables through PID and RegimenID.

Column	Data type	Distinct	Missing	Description
PID	unique id	31,421	0	> PID in Admissions table
DrugID	int	706,670	0	unique
RegimenID	int	59,639	32,683	High missing values
AgeWhenDrugGiven	int	34,184	0	2 – 36,720
DrugLabel	varchar(max)	301	0	Count:[1- 81,501]
DrugDose	float	435	3	Explaining DrugLabel
DoseUnit	varchar(max)	12	3	Explaining DrugDose
DrugLabelMapped	varchar(max)	291	11,971	Mapped 1-1 to DrugLabel
DrugTherapyType	varchar(max)	19	11,971	High missing values
Number of rows		706,670		

Notes:

(1) Some missing RegimenID can be replaced by:

a. referencing to ChemoRegimen table, when AgeAtRegimenStartDate = AgeWhenDrugGiven, leaving with 9,799 missing values

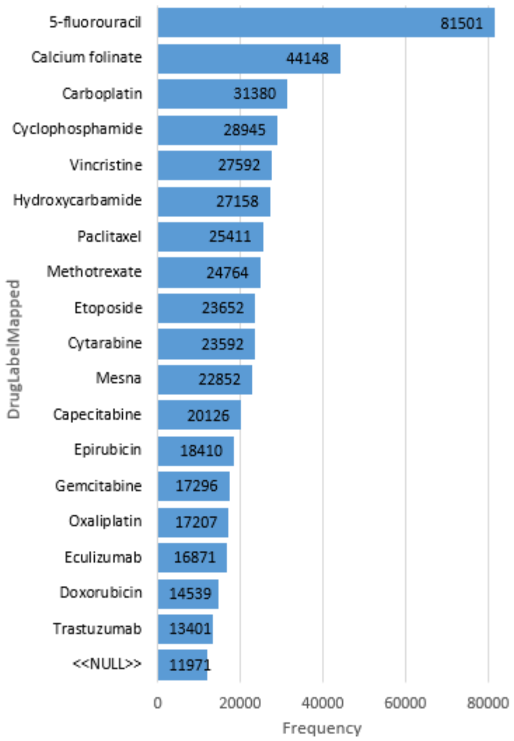
b. referencing to ChemoCycles table, when AgeWhenCycleStarted = AgeWhenDrugGiven, leaving with 5,151 missing values

(3) There are 11 DrugLabel that are not mapped to DrugLabelMapped and DrugTherapyType

(4) Potential activities: DrugLabel / DrugLabelMapped

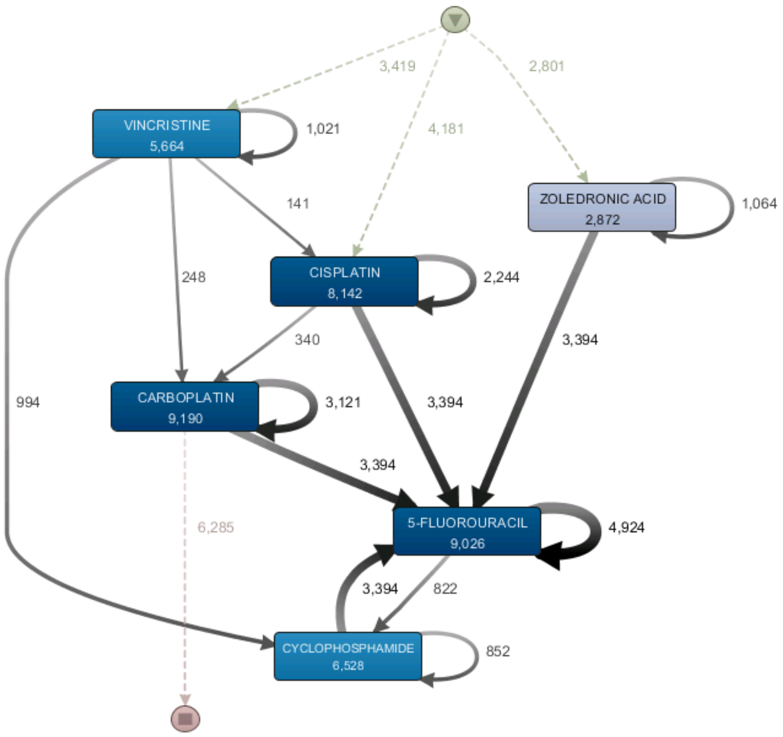
(5) Potential splitting attributes: DrugTherapyType

Attribute values:



DrugTherapyType	freq
Cytotoxic	515677
Calcium folinate	44148
Biological Modifier	37140
Monoclonal Antibody	25514
Cytoprotective	22852
<<NULL>>	11971
Other	11676
Biphosphonate	10380
Glucocoeticoid	9628
Kinase Inhibitor	7536
Hormone therapy	5136
Vaccine	2859
Anti-emetic	1183
Bisphosphonate	566
Hormone	201
Additive	142
Antibiotic	40
Antiviral	18
Cytokine Growth Factor	3

Fuzzy model:



(1.7% activities, 8.4% paths, showing case frequency)

Details:

Events	: 706,670
Cases	: 76,995
Events per case	: min 1, mean 9, max 448
Activities	: 301
Mean case duration	: 81.7 days
Median case duration	: 27 days
Case duration	: 0 days - 17 years, 2 months, 13 days
Variants	: 12,791

5) ChemoRegimens

The ChemoRegimens table provides detail information regarding chemotherapy regimens of the patient. This is linked to other tables through PID and RegimenID.

Column	Data type	Distinct	Missing	Description
PID	unique id	31,354	0	Count:[1-33]
RegimenID	int	67,707	0	unique
AgeAtRegimenStartDate	int	22,744	0	Relative
RegimenLabel	varchar(max)	2,945	0	Count:[1-1389]
Intent	varchar(max)	7	12,687	
RegimenLabelMapped	varchar(max)	547	8,972	
LinkedDiagnosisId	int	32,490	4,360	
LinkedDiagnosisICD	varchar(max)	68	4,360	
DiagnosisLinkMethod	varchar(max)	3	4,360	
CourseOfSameChemo	int	14	0	
Number of rows		67,707		

Notes:

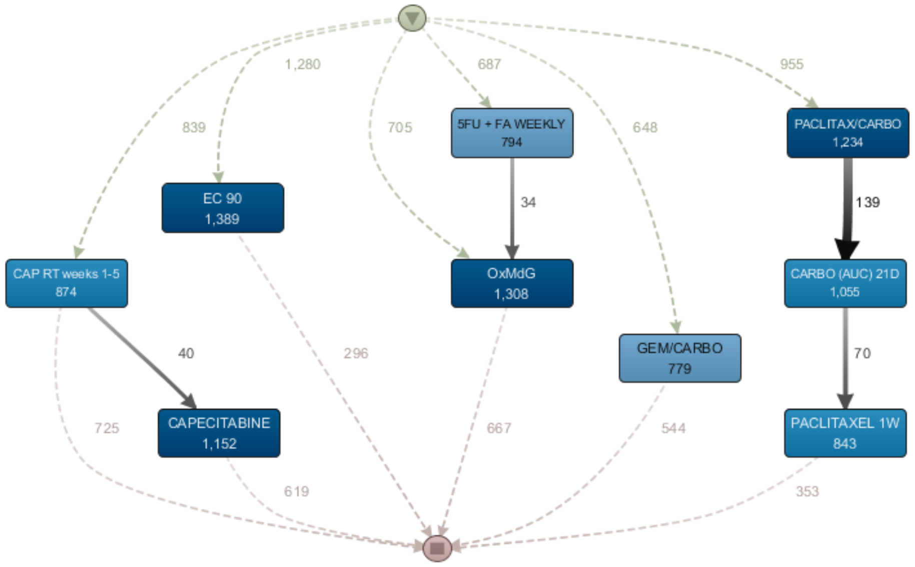
(1) Potential activities: RegimenLabel,

(2) Potential splitting attributes: Intent, LinkedDiagnosisICD, DiagnosisLinkMethod.

Details:

Events	: 67,707
Cases	: 31,354
Events per case	: min 1, mean 2, max 33
Activities	: 2,951
Mean case duration	: 37.7 weeks
Median case duration	: 0 milliseconds
Case duration	: 0 days – 18 Years, 5 Months, 24 Days
Variants	: 2,951

Fuzzy model:



(0.3% activities, 0% paths) (showing case frequency)

The above fuzzy model shows a model where a case consists only of a small number of activities. Based on the following details, each case consists of 1 to 33 activities, with an average of 2.

6) Death

The Death table provides information about the AgeAtDeath and LastCycleBefore Death. It is linked to other tables through PID.

Column	Data type	Distinct	Missing	Description
PID	unique id	17,701	0	
AgeAtDeath	int	11,214	0	28 - 36,649
LastCycleBeforeDeath	int	10,504	1,317	(in age- days) 2 – 36,089
Number of rows		17,701		

Notes:

(1) Potential activity: Death

(2) There is no attributes potentially useful as splitting attribute in this table.

(3) Because this table consists of only one activity, it wouldn't be useful to use this table alone in process mining, and will be combined with the other tables.

7) Diagnosis

The Diagnosis table provides information about the patient diagnosis. It is linked to other tables through PID.

Notes:

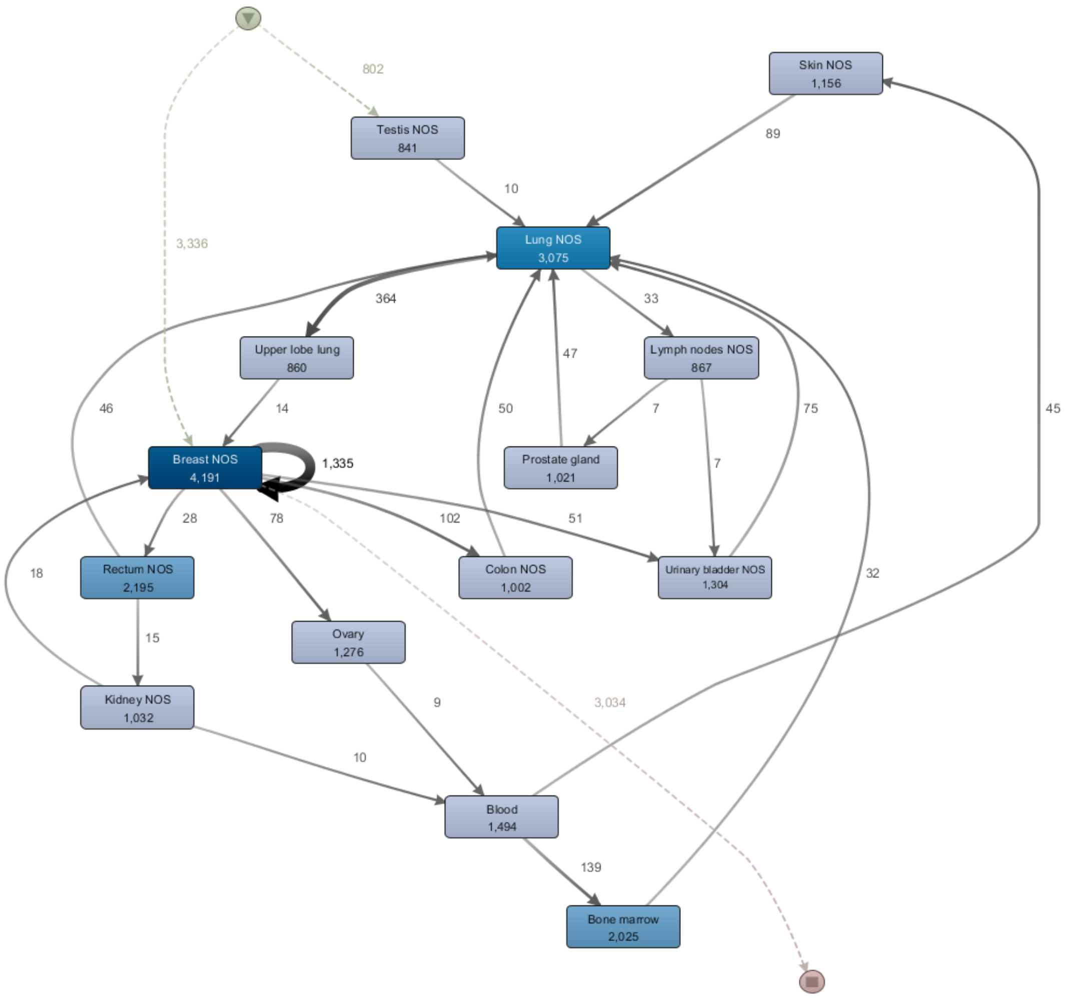
- Potential activities: dx_ICD10Label.

Column	Data type	Distinct	Missing	Description
PID	unique id	30,753	0	
AgeAtDiagnosis	int	20,132	0	0 - 36609
YearOfDiagnosis	int	63	0	1921-2015
dx_DiagnosisID	int	56,123	0	unique
dx_ICD10CDS	nvarchar(50)	507	7,892	
dx_ICD10Label	nvarchar(100)	518	6,422	
TStage	nvarchar(100)	32	44,387	
NStage	nvarchar(100)	19	44,905	
MStage	nvarchar(100)	12	45,634	
StageLabel	nvarchar(100)	74	45,147	
dx_DiseasePhase	int	3	421	
dx_DiseasePhaseLabel	nvarchar(100)	3	421	
dx_CancerStatus	int	7	77	
dx_CancerStatusLabel	nvarchar(100)	7	77	
dx_MorphologyCDS	nvarchar(50)	471	8,464	
dx_MorphologyCode	int	521	8,261	
dx_MorphologyLabel	nvarchar(100)	521	8,261	
dx_SiteCDS	nvarchar(50)	265	614	
dx_SiteLabel	nvarchar(100)	278	0	
dx_Her2Status	int	8	54,774	
HER2Status_Label	varchar(35)	6	54,784	
dx_EstrogenReceptorStatus	int	11	54,540	
OestrogenReceptorStatus_Label	varchar(16)	10	54,545	
dx_ProgesteroneReceptorStatus	int	10	54,762	
ProgesteroneReceptorStatus_Label	varchar(16)	9	54,775	
DistrictLevelPostcodeAtDiagnosis (if known)	varchar(8000)	166	51,108	
DrivingDistanceFromLTHT_Miles	float	1,065	54,570	
Number of rows		56,123		

Details:

Events	: 56,123
Cases	: 30,753
Events per case	: min 1, mean 2, max 21
Activities	: 278
Mean case duration	: 22.9 months
Median case duration	: 0 millis
Case duration	: 0 days – 18 Years, 5 Months, 24 Days
Variants	: 6,063

Fuzzy model:



(4.8% activities, 0% paths, showing case frequency)

8) Outpatients

The Outpatients table provides information about the outpatient visits. It is linked to other tables through PID.

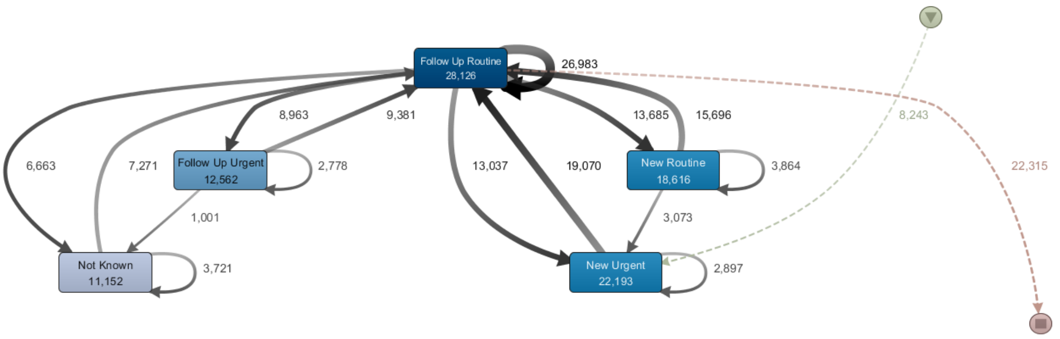
Column	Data type	Distinct	Missing	Description
PID	unique id	28,878	0	
AgeAtTimeOfOPClinic	int	34,438	0	
op_AppointmentTypeCode	int	166	28,580	
AppointmentTypeDescription	nvarchar(255)	69	47,564	NULL='Not Known'
Number of rows		973,177		

Notes:

(1) Potential activities: AppointmentType Description ('Not Known' treated as an activity).

(2) Potential splitting attribute: none.

Fuzzy model:



(7% activities, 7% paths, showing case frequency)

Details:

Events	: 926,797
Cases	: 28,878
Events per case	: min 1, mean 32, max 399
Activities	: 69
Mean case duration	: 54.7 months
Median case duration	: 42.8 months
Case duration	: 0 days – 14 Years, 6 Months, 19 Days
Variants	: 23,868

9) Patients

The Patients table provides information about the outpatient visits. It is linked to other tables through PID.

Column	Data type	Distinct	Missing	Desc
PID	unique id	31,511	0	
Gender	nvarchar(100)	2	27	
EthnicCategory	nvarchar(100)	20	1514	
AgeAtDeath	int	11,214	13,810	
DistrictLevelPostcode	nvarchar(4)	693	45	
DrivingDistanceFromLTHT_Miles	float	4,341	3,655	
DayOfWeekOfBirth	nvarchar(30)	7	1	
Number of rows		31,511		

Notes: (1) Potential activities: death (duplicate with Death table). (2) Missing ethnic categories: NULL=1514, Not given – 4881, Not collected – 234, Not stated – 192. (3) Potential splitting attributes: Gender, EthnicCategory, DistrictLevelPostcode, and DistrictLevelPostcode. (4) The Patients table is potentially useful for patient characterisation, such as proportion of male versus female patients and proportion of patient ethnicity.

10) Radiotherapy

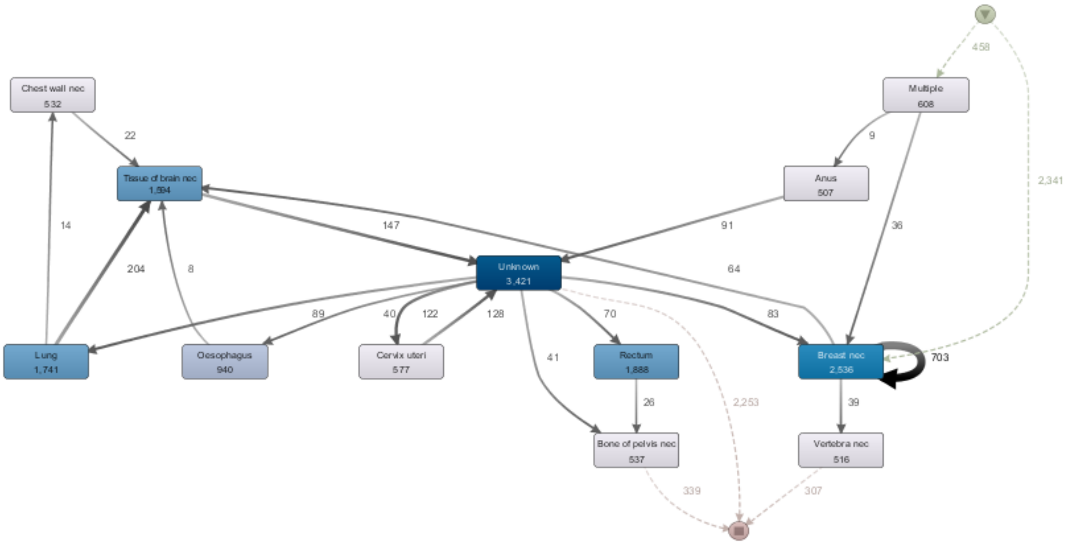
The Radiotherapy table provides information about the radiotherapy treatment. It is linked to other tables through PID. Radiotherapy is one of three possible treatment for cancer, along with chemotherapy and surgery.

Column	Data type	Distinct	Missing	Description
PID	unique id	16,792	0	
AgeAtRadiotherapy	int	15,212	0	one invalid (-13959)
IntentCode	nvarchar(50)	4	4,495	
IntentLabel	nvarchar(100)	5	3,298	
SiteCode	nvarchar(50)	232	5,361	
SiteLabel	nvarchar(100)	236	4,720	
Number of rows		31,703		

Notes: (1) Potential activity: 'Radiotherapy',

(2) Potential splitting attributes: IntentLabel, SiteLabel

Fuzzy model:



(4.7% activities, 0% paths) (showing case frequency)

The above fuzzy model shows a possibility to use SiteLabel as the activity name. It resulted in a model of site progression of cancer. The model shows that the most frequent activity is 'Unknown', which should be excluded from the model, with a risk of losing information of the most common site label.

Details:

Cases	: 16,792
Events	: 31,160
Events per case	: min 1, mean 2, max 30
Mean case duration	: 45.3 weeks
Case duration	: 0 millis – 12 years 7 months 6 days
Variants	: 3,361

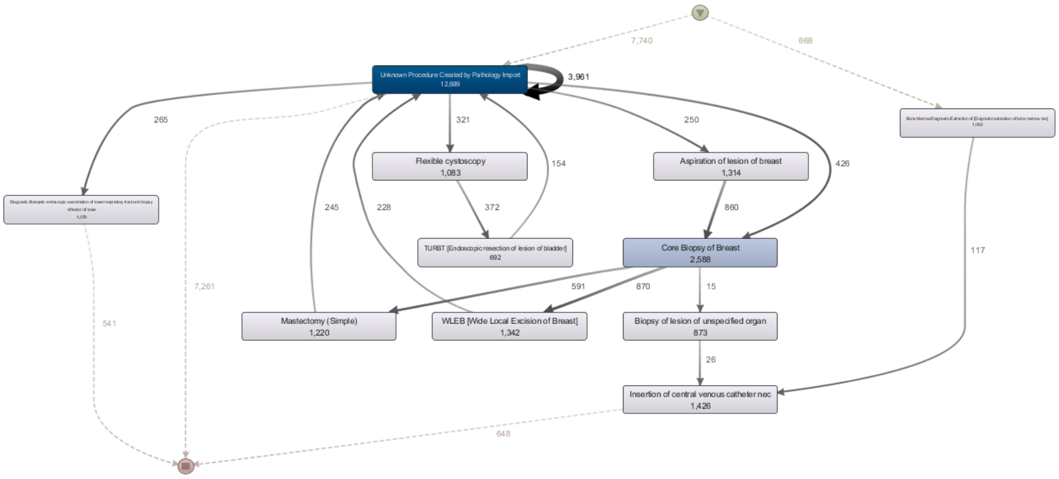
11) Surgery

The Surgery table provides information about the surgery treatment. It is linked to other tables through PID. Surgery is one of three possible treatment for cancer, along with chemotherapy and radiotherapy.

Column	Data type	Distinct	Missing	Description
PID	unique id	21,395	0	Case id
AgeAtSurgery	int	22,199	0	Activity
ProcedureCode	nvarchar(50)	1,237	26,134	
ProcedureLabel	nvarchar(100)	1,345	6	23,389 Unknown
Number of rows		74,889		

Notes: Potential activity: 'Surgery' / ProcedureLabel

Fuzzy model:


7,740
668
Unknown Procedure Created by Pathology Import
12,689
3,961
265
321
250
Bone Marrow Diagnostic Extraction of [Diagnostic extraction of bone marrow nec]
1,092
Flexible cystoscopy
1,083
154
Aspiration of lesion of breast
1,314
426
Diagnostic fibreoptic endoscopic examination of lower respiratory tract and biopsy of lesion of lower
1,079
245
228
372
860
TURBT [Endoscopic resection of lesion of bladder]
692
Core Biopsy of Breast
2,588
117
7,261
591
870
15
541
Mastectomy (Simple)
1,220
WLEB [Wide Local Excision of Breast]
1,342
Biopsy of lesion of unspecified organ
873
26
Insertion of central venous catheter nec
1,426
648

(1% activities, 0% paths, showing case frequency)

The above fuzzy model shows that the most frequent activity is 'Unknown procedure created by Pathology import'. A discussion with clinical expert confirmed that this activity can be further analysed from the Pathology table, which unfortunately was not included in the PPM Chemotherapy dataset.

Details:

Cases	: 21,395
Events	: 74,695
Events per case	: min 1, mean 3, max 50
Mean case duration	: 22.5 months
Median case duration	: 22.6 weeks
Case duration	: 0 millis – 12 years 7 months 6 days
Variants	: 3,361

12) TestResultsBlood

The TestResultsBoold table provides information about the results of blood tests. It is linked to other tables through PID.

Column	Data type	Distinct	Missing	Description
PID	unique id	29,151	0	KTPId
AgeAtOrderDate	int	34,971	0	0 – 36,751
Value	varchar(50)	21,063	577	Numbers, some started with < or >
Units	varchar(50)	16	1,037,211	
Term	varchar(256)	34	0	
Number of rows		17,814,931		

Notes: (1) Potential activities: Term; (2) Potential splitting attribute: Term+Value

(3) A possible data quality issue is that some values started with < or > when they are expected to be in a numeric format.

(4) TestResultsBlood table contains most of the data in the PPM Chemotherapy dataset (71 million/ 83%)

13) TestResultsMicrobiology

The TestResultsMicrobiology table provides information about the results of microbiology tests. It is linked to other tables through PID.

Column	Data type	Distinct	Missing	Description
PID	unique id	15,678	0	KTPId
AgeAtOrderDate	int	27,823	0	
Source	varchar(max)	23	0	
Positive	bit	2	0	0/1
PossibleContaminant	bit	2	0	0/1
Number of rows		118,947		

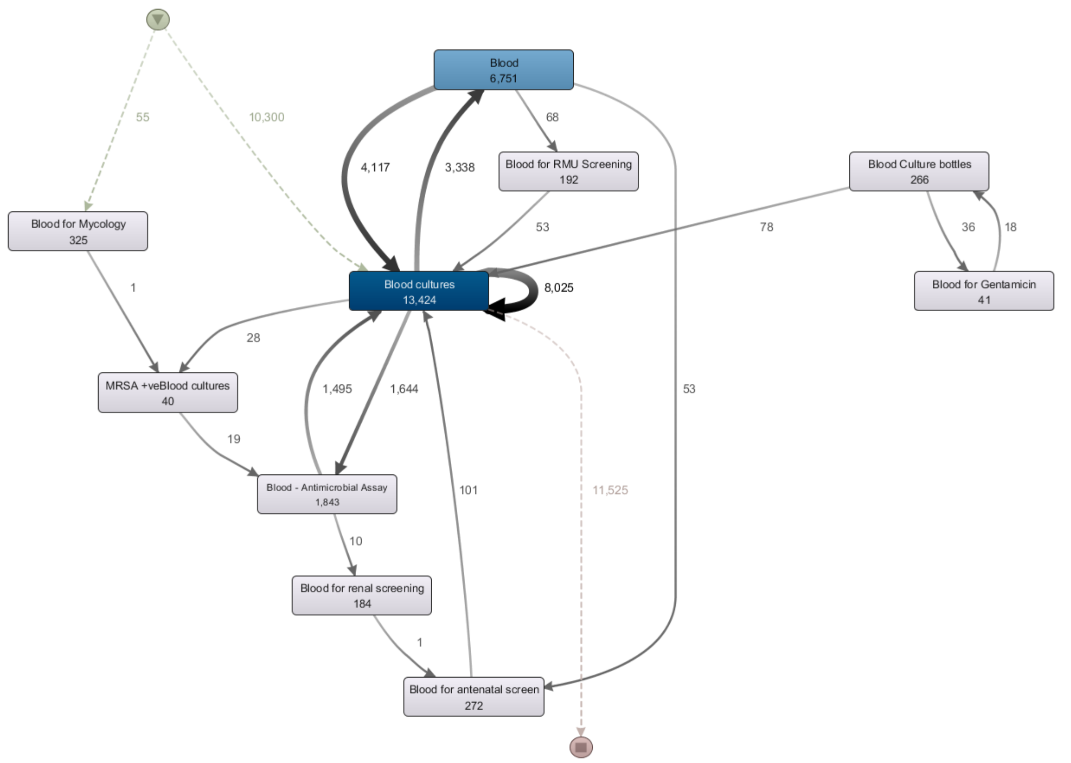
Notes: (1) Possible activities: Source

(2) Possible splitting attributes: Source+Positive

Details:

Cases	: 15,678
Events	: 88,049
Events per case	: min 1, mean 6, max 187
Mean case duration	: 12.9 months
Median case duration	: 29 days
Case duration	: 0 millis – 11 years 9 months 28 days
Variants	: 3,093

Fuzzy model:



(43% activities, 0% paths) (showing case frequency)

Conclusion

This experiment shows that all tables except Patients table are potentially useful for process mining. Patients table contains no timestamped event, but is useful for patient characterisation. Some tables can be used for process mining combined with other tables (Death, Radiotherapy, Surgery), some tables cannot be used as event log but will be useful as a reference table (Patients), while the other 9 tables can analysed using process mining by themselves. Some data quality issues are identified and has been described for each table in this experiment.

D.2 Process mining of chemotherapy pathways

<table>
<tr><td rowspan="3">UNIVERSITY OF LEEDS
School of Computing</td><td><b>EXPERIMENT DOCUMENTATION</b></td><td rowspan="2"><b>Date of experiment</b>
05/10/2018</td></tr>
<tr><td><b>Experiment title:</b>
Process mining of EC-90 chemotherapy pathways of breast cancer patients</td></tr>
<tr><td><b>Researcher's name:</b>
Angelina Kurniati</td><td><b>Experiment code</b>
PPMC-SBRI-BC001</td></tr>
</table>

Area of investigation

This experiment is for using process mining to explore variations in chemotherapy pathways for breast cancer patients.

Data source

The dataset is a subset of SBRI dataset from PPM Chemotherapy data (712 variants of 738 patients).

Research question

Can process mining be used to explore variations in chemotherapy pathways for breast cancer patients?

Hypothesis

Process mining can be used to explore variations in chemotherapy pathways for breast cancer patients.

Method

1) **Extraction and transformation** has been done in CHE-ALL000. Patients were included if they had a diagnosis of metastatic breast cancer (ICD-10 C50) and received adjuvant epirubicin and cyclophosphamide (EC-90) chemotherapy.

3) **Create** event log of all events with [case_id, activity, timestamp, tsource] format, with:

case_id = PID/ KTPId

activity = {identified in CHE-ALL000}

timestamp = {identified in CHE-ALL000}

2) Save as .csv files

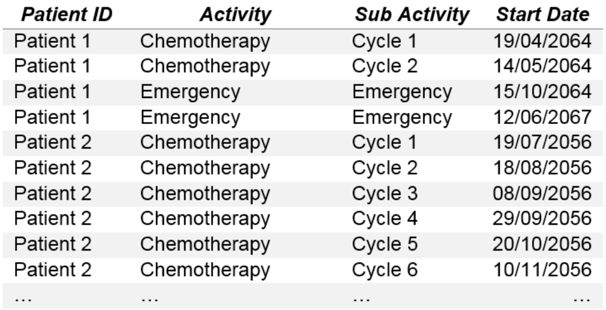
3) **Load** into DISCO and ProM (if possible), and analyse the result.

Results and Discussion

1) Extraction and transformation has been done by selecting events of interest.

Event name	Occurence	Percentage
Neutropenic	836	17.2%
Cycle 1	732	15.0%
Cycle 2	725	14.8%
Cycle 3	699	14.3%
Emergency	611	12.5%
Cycle 4	487	9.9%
Cycle 5	402	8.3%
Cycle 6	380	7.8%

2) The event log is illustrated below.



Patient ID	Activity	Sub Activity	Start Date
Patient 1	Chemotherapy	Cycle 1	19/04/2064
Patient 1	Chemotherapy	Cycle 2	14/05/2064
Patient 1	Emergency	Emergency	15/10/2064
Patient 1	Emergency	Emergency	12/06/2067
Patient 2	Chemotherapy	Cycle 1	19/07/2056
Patient 2	Chemotherapy	Cycle 2	18/08/2056
Patient 2	Chemotherapy	Cycle 3	08/09/2056
Patient 2	Chemotherapy	Cycle 4	29/09/2056
Patient 2	Chemotherapy	Cycle 5	20/10/2056
Patient 2	Chemotherapy	Cycle 6	10/11/2056
…	…	…	…

3) The event log above has been loaded to ProM and DISCO for further analysis. The resulted process model, trace variant diagram, and dotted chart presented in Figure 5.3 – Figure 5.5 in Section 5.2.3.

4) Further analysis has been done to examine the cycles leading to an emergency admission or a neutropenic condition. The following table shows that most patients who had emergency admission got it after Cycle 3, Cycle 6, or Cycle 1; while most patients who had Neutropenic condition got it after Cycle 3, Cycle 2, or Cycle 1.

Cycle number	Cycle leads to Emergency		Cycle leads to Neutropenic	
	N (%)	med; mean	N (%)	med; mean
Cycle 1	81 (11)	8 d; 18.4 d	94 (13)	19 d; 23.1 d
Cycle 2	52 (7)	8 d; 43.9 d	123 (17)	19 d; 20.6 d
Cycle 3	117 (16)	28 d; 27.3 w	142 (19)	18 d; 61.1 d
Cycle 4	64 (9)	14d; 27.3 w	84 (11)	19 d; 16 d
Cycle 5	22 (3)	13.5 d; 19.2 w	70 (9)	19 d; 33.5 d
Cycle 6	90 (12)	13.5 m; 20.8 m	57 (8)	14 d; 26.4 w

5) Another discussion with clinical experts was that it seems that a lot of patients stopped chemotherapy after Cycle 3. This is presented in Figure 5.4 as the second (n=56; 7.6%) and third trace variant (n=37; 5%). The clinical expert explained that this was because the analysis was done in only one regimen (EC-90). The regimen might have changed in the subsequent cycles. This condition is out of the scope of the experiment.

Conclusion

This experiment shows that process mining of routine data can show extensive variations from standard chemotherapy pathways including incomplete treatment and adverse events. Future work is needed to explore potential causal links and understand changes in the pathways over time.

Appendix E
Analysis of the PPM Cancer dataset

E.1 Table description of the PPM Cancer dataset

The following table contains list of table and the details of PPM Cancer dataset.

#	table name	# columns	data type*			
			id	char	num	time
1	Admissions	63	1	36	22	4
2	Annotations	71	1	29	36	5
3	AssessmentActivities	8	2	0	2	4
4	AssessmentResults	5	2	1	2	**0**
5	Assessments	3	2	0	1	**0**
6	COSDInvestigations	24	0	14	6	4
7	COSDPathology	39	0	25	9	5
8	COSDRegistration	84	0	68	9	7
9	COSDSurgery	32	0	18	8	6
10	CWTReferralTreatmentComplete	52	0	26	22	4
11	CancerWaitingTimes	142	1	58	62	21
12	CareEpisodes	56	1	30	19	6
13	ChemoCycles	94	1	43	42	8
14	ChemoDrugs	153	1	75	67	10
15	ChemoRegimens	77	1	34	35	7
16	Consultations	77	1	38	35	3
17	DataForIBMPOC	4	0	3	1	**0**
18	Diagnosis	318	1	153	155	9
19	Event	23	1	2	15	5
20	EventAdmission	13	1	0	12	**0**
21	EventBloodTest	59	1	1	56	1
22	EventMetadata	69	5	12	41	11
23	EventMetadataSummaries	6	1	2	2	1
24	EventWardStay	22	1	16	5	**0**
25	InvestigationEx	50	1	20	26	3
26	Investigations	85	1	47	32	5
27	LastContacts	55	0	35	18	2
28	MDTReview	82	1	45	29	7
29	Organisation	10	1	6	3	6
30	Outpatients	99	1	52	40	3
31	Pathology	335	1	160	171	3
32	PatientOrganisation	8	1	4	3	**0**
33	PatientOrganisationVersion	37	1	25	8	3
34	PatientPCT	3	0	2	1	**0**
35	Patients	90	1	55	29	5
36	Radiotherapy	103	1	47	48	7
37	RadiotherapyEx	76	1	33	37	5
38	Referrals	99	1	53	36	9
39	StratifiedMedicineResultsBase	10	0	7	3	**0**
40	StratifiedMedicine_TabulatedResultsXML	233	0	203	3	27

#	table name	# columns	data type*			
			id	char	num	time
41	Surgery	160	1	97	55	7
42	TrialActivityContactOrg	105	1	56	35	14
43	TrialConsultationActions	22	1	9	10	3
44	TrialRecruitmentInstitution	57	1	20	28	8
45	TwoWeekWaitReferrals	25	1	17	6	2
46	WardStays	81	1	45	29	6
47	Watch	2	1	2	2	2
48	WatchDefinition	12	1	97	8	7
Total		3,333	40	1,724	1,324	245

*****id*** *includes uniqueidentifier*

*****char*** *includes char, nchar, nvarchar, varchar*

*****num*** *includes bit, float, int, numeric, real, tinyint, varbinary*

*****time*** *includes date, datetime*

E.2 Data selection of the PPM Cancer dataset

The following table contains description of table name, columns, and data types in the database which was used in this study.

Code: C=case_id, A=activity_id, R=resource, T=timestamp

#) TABLE_NAME	COLUMN_NAME	DATA_TYPE	CODE
1) Admissions	em_AdmissionDate	datetime	T, A
	em_ContactSpecialityLabel	nvarchar	R
	em_DischargeDate	datetime	T
	em_DischargeMethodLabel	nvarchar	A
	em_PatientID	int	C
2) ChemoCycles	ecc_CycleContactSpecialityLabel	nvarchar	R
	ecc_CycleNumber	int	A
	Ecc.PatientID	int	C
	ecc_CycleStartDate	datetime	T
3) ChemoDrugs	ecd_DrugEndDate	datetime	T
	ecd_DrugLabel	nvarchar	A
	ecd_DrugStartDate	datetime	T
	Ecd_CycleCOntactSpecLabel	nvarchart	R
	ecd_PatientID	int	C
4) ChemoRegimens	ec_ContactSpecialityLabel	nvarchar	R
	ec_RegimenEndDate	datetime	T
	ec_RegimenLabel	nvarchar	A
	ec_RegimenStartDate	datetime	T
	ec_PatientID	int	C
5) Consultations	eb_ConsultationDate	datetime	T
	eb_ContactMethodLabel	nvarchar	A
	eb_ContactSpecialityLabel	nvarchar	R
	eb_PatientID	int	C
6) Investigations	en_PatientID	int	C
	en_InvestigationLabel	nvarchar	A
	en_ContactSpecLabel	nvarchar	R

#) TABLE_NAME	COLUMN_NAME	DATA_TYPE	CODE
	en_ContactTypeLabel	int	C
7) Diagnosis	dx_ContactSpecLabel	nvarchar	R
	dx_DiagnosisDate	datetime	T
	dx_ICD10CDS3	nvarchar	A
	dx_ContactTypeLabel	nvarchar	R
	dx_PatientID	int	C
8) MDTReview	ev_EventDate	datetime	T
	ev_EventDetail	nvarchar	A
	ev_PatientID	int	C
	ev_TeamName	nvarchar	R
9) Outpatients	op_AppointmentDate	datetime	T
	op_AppointmentTypeCodeLabel	nvarchar	A
	op_ClinicConsultantContactSpecialityLabel	nvarchar	R
	op_ClinicDate	datetime	T
	op_ClinicType	nvarchar	A
	op_PatientID	int	C
10) Pathology	esp_ContactSpecialityLabel	nvarchar	R
	esp_HistologyReportDate	datetime	T
	esp_PathologyDate	datetime	T
	esp_PatientID	int	C
	esp_ReceivedDate	datetime	T
	esp_SiteCodeLabel	nvarchar	A
	esp_SpecimenTypeLabel	nvarchar	A
11) Patients	pt_BirthDate	datetime	T
	pt_CauseOfDeathCode	int	A
	Pt_DeathPlace	nvarchar	A
	pt_DeathDate	datetime	T
12) Radiotherapy	er_BookedDate	datetime	T
	er_ContactTypeLabel	nvarchar	R
	er_EndDate	datetime	T
	er_EventDate	datetime	T
	er_PatientID	int	C
	er_TypeCodeLabel	nvarchar	A
13) Referrals	ef_AcceptedDate	datetime	T
	ef_FirstAppointmentDate	datetime	T
	ef_PatientID	int	C
	ef_ReceivedDate	datetime	T
	ef_ReferralDecisionDate	datetime	T
	ef_ReferralTypeLabel	nvarchar	A
	ef_ReferredByContactSpecLabel	nvarchar	R
	ef_ReferredToContactSpecLabel	nvarchar	R
14) Surgery	es_DecisionDate	datetime	T
	es_MainProcedureLabel	nvarchar	A
	es_MethodLabel	nvarchar	A
	es_PatientID	int	C
	es_ProcedureText	nvarchar	A
	es_SurgeonSpecialityLabel	nvarchar	R
	es_SurgeryDate	datetime	T

E.3 Analysing PPM Cancer data change points

UNIVERSITY OF LEEDS School of Computing	**EXPERIMENT DOCUMENTATION**	**Date of experiment** 25/06/2019
	Experiment title: PPM Cancer data change points	
	Researcher's name: Angelina Kurniati	**Experiment code** PPMQ-CH0002

Area of investigation

This experiment is the detection of the change points of the data. The findings in the previous experiment show a change pattern in the system usage. This experiment focused on finding change points where the system has been potentially changed.

Data source

The dataset is all events of cancer patients in the PPM Cancer data, which was accessed in the secure room in LIDA. Specifically, the data of monthly records is analysed. Details of the R libraries used in this experiment is in Appendix F.2.

Research question

Is it possible to detect change points where system is potentially changed?

Hypothesis

It is possible to detect change points based on the system usage over time.

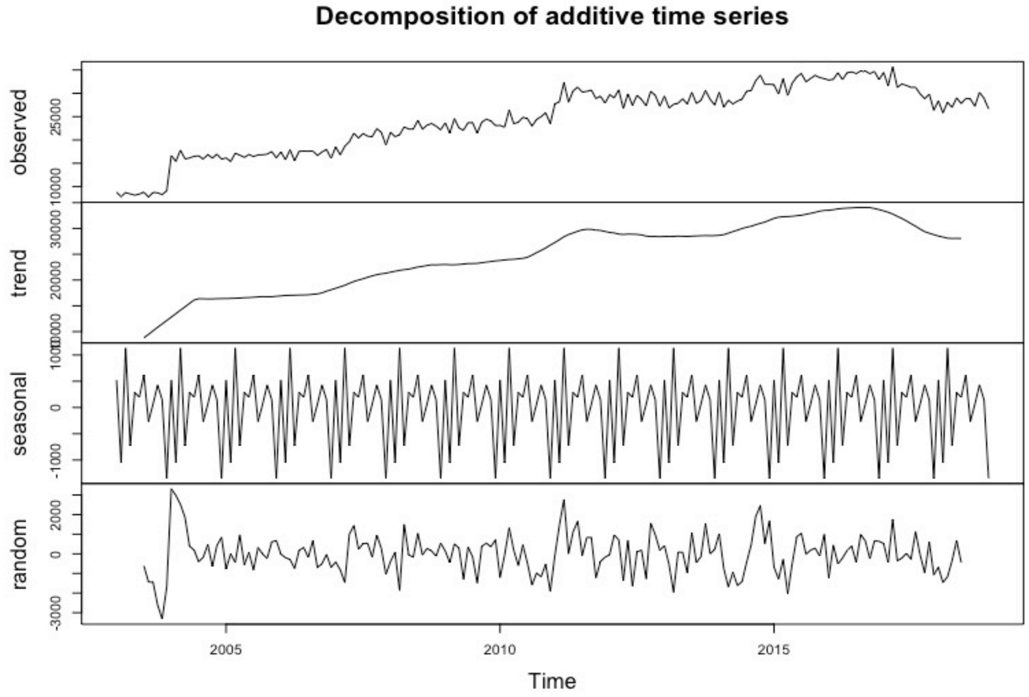
Method

For each event:

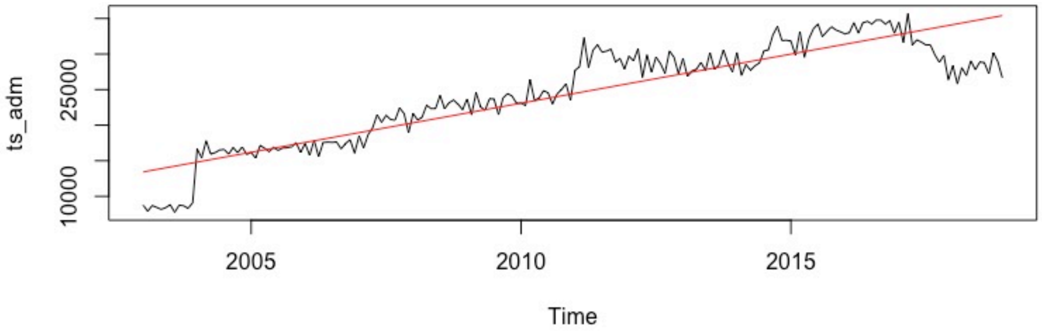
1) Query to calculate monthly frequency of each activity from 2003 to 2018.
2) Create a time-series (***ts)*** object using R library ***stats.*** Frequency = 12.
3) Signal decomposition to subtract trend and seasonal pattern from the observed pattern. Time series patterns are trend (based on 12-moving average), seasonal (monthly average), and random. R library used in this experiment is ***fpp2***.
4) Fit a linear model to analyse the trend, using R library ***stats.*** This is analysed based on the coefficient, adjusted R^2, and p-value. The coefficient represents the intercept and slope in the linear model. The adjusted R^2 shows how well the model is fitting the actual data. The p-value shows the significance of the relationship.
5) SPC chart to detect time points with significant change of activity frequency over time. R library used in this experiment is ***qicharts2***.
6) Analyse the results.

Results

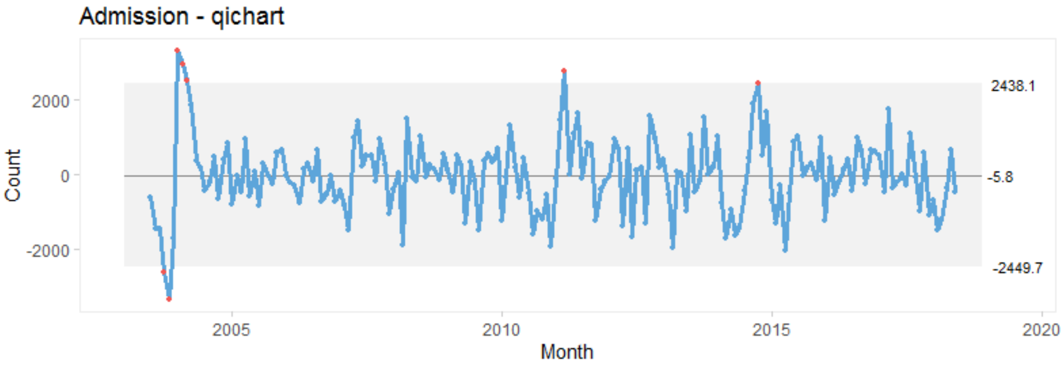
Admission *(Min 7761; Med 26102; Mean 24467; Max 35669)*



Fitted trend: *(coef = 115.0181, adjusted* R^2 *= 0.8144, p-value < 2.2e-16)*



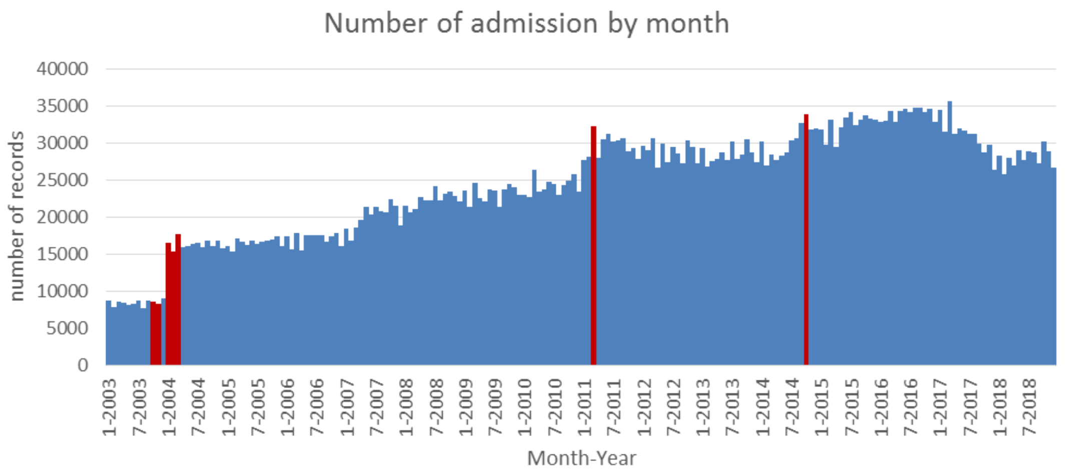
SPC to detect change points:



Change points:

Month	10-2003	11-2003	1-2004	2-2004	3-2004	3-2011	10-2014
value	-2600.5	-3323.6	3316.4	2979.1	2522.8	2771.4	2456.9

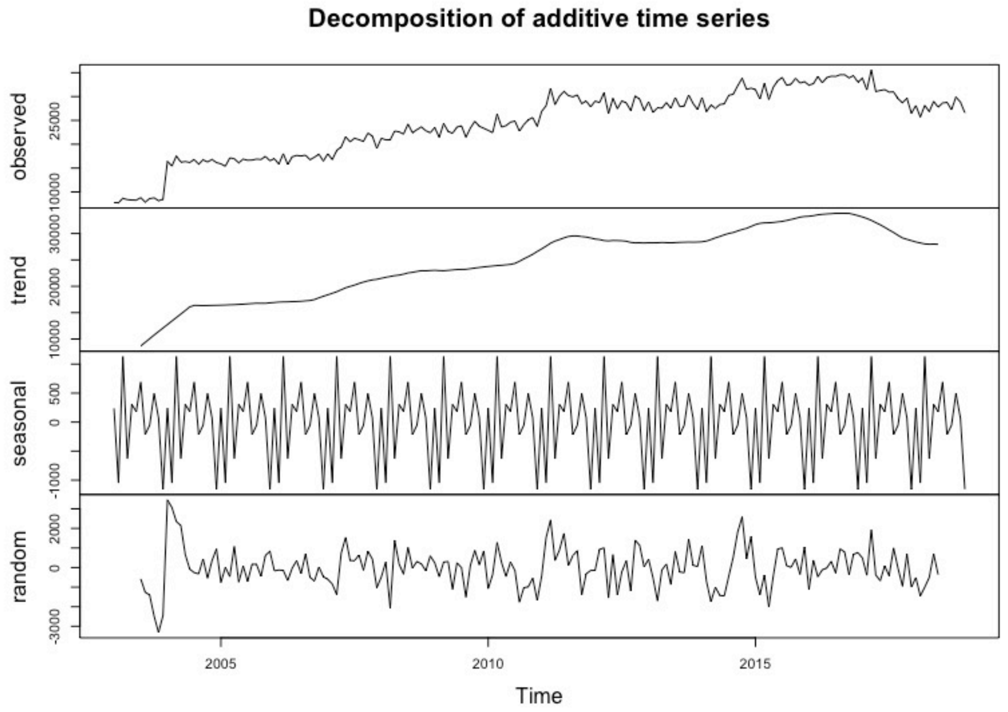
Change points mapped into bar chart of monthly records:



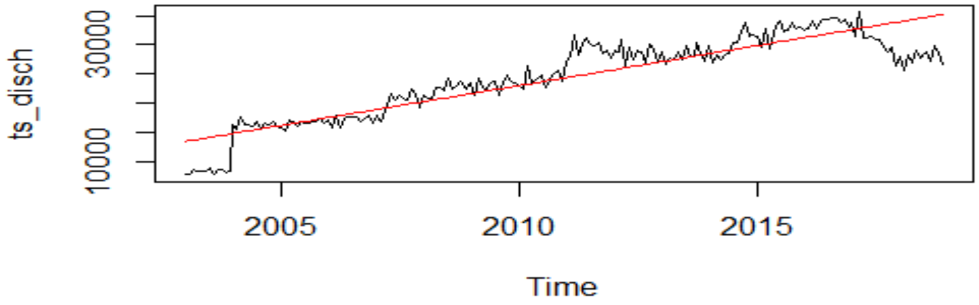
The trend of *Admission* shows an increasing pattern by 115 records per month and is higher than the trend of *Diagnosis*. The change points detected in October-November 2003, January – March 2004, March 2011, and October 2014.

The changes in October – November 2003 reflected the early stage of PPM EHR system implementation. The change in March 2011 and Oct 2014 were likely related to the major changes in the PPM EHR system. In 2011, PPM EHR system were started to be adopted to manage data in the whole hospital, where previously used only to record events related to cancer treatment. In 2014, PPM EHR system has been improved to join the Leeds Care Records (LCR) and connected to other providers such as General Practitioners (GPs), community and adult social care.

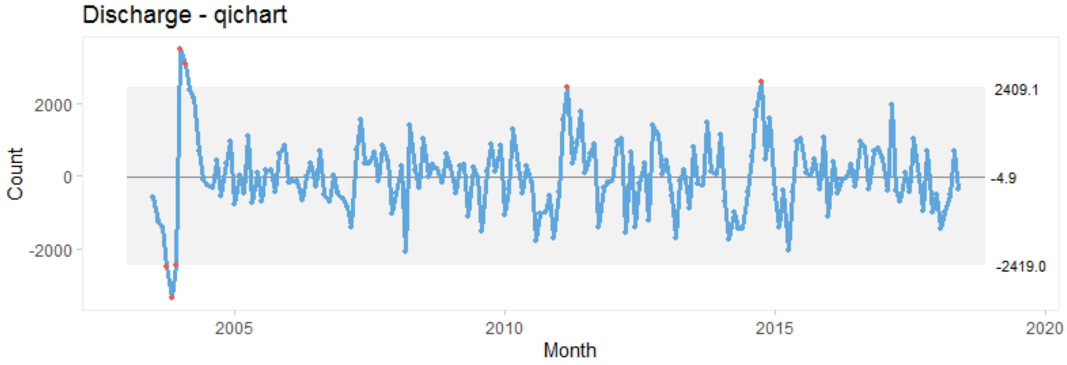
***Discharge** (Min 7754; Med 26025; Mean 24338; Max 35564)*



Fitted trend: *(coef = 113.6, adjusted R^2 = 0.8146, p-value < 2.2e-16)*



SPC to detect change points:



Change points:

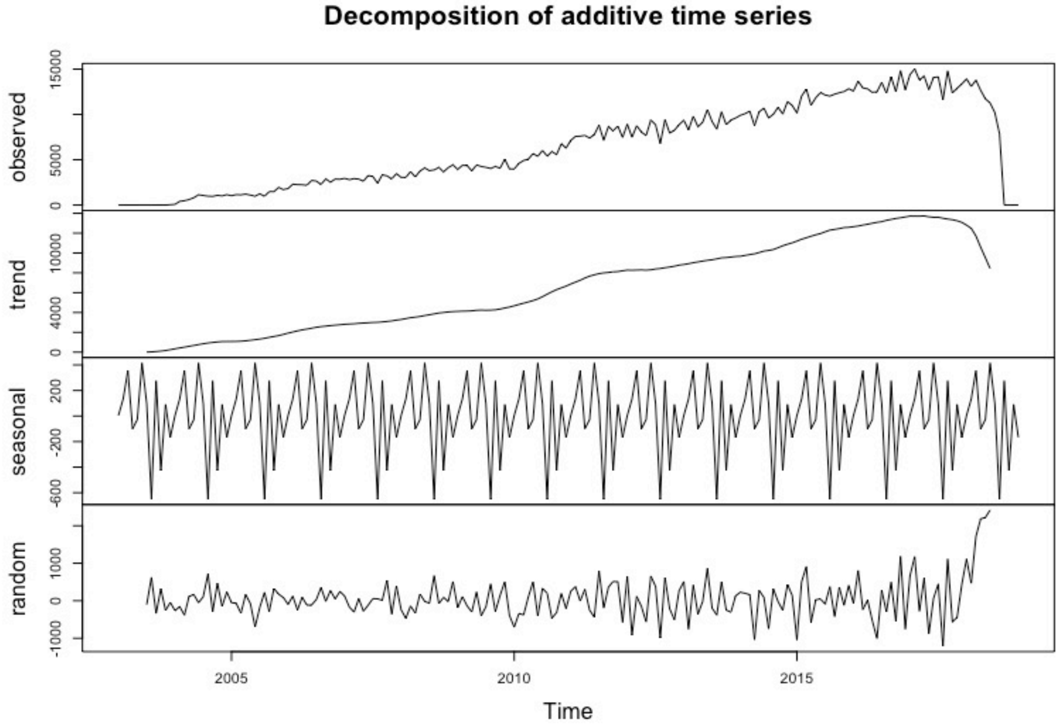
Month	10-2003	11-2003	12-2003	1-2004	2-2004	3-2011	10-2014
value	-2474.9	-3317.6	-2448.4	3456.5	3063.8	2431.5	2583.8

The number of *Discharge* records by month is very similar to those of admissions.

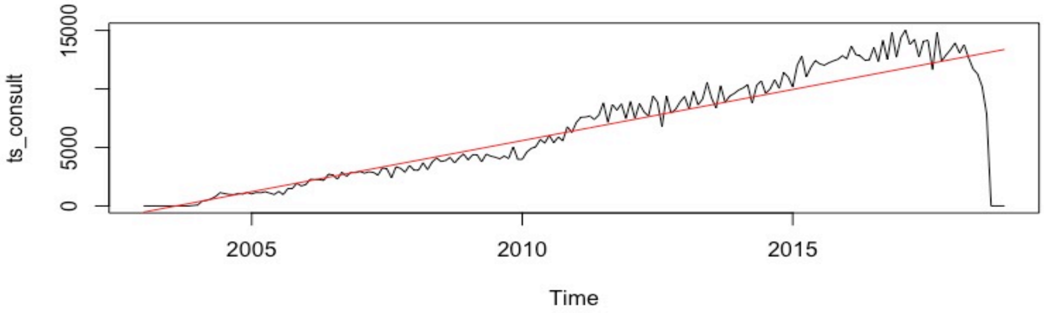
The trend is that there is an increase of 113.6 per month and is higher than the trend of *Diagnosis*. The trend shows that *Discharge* records increased by 113.66 per month. Three periods where change potentially happened are in October 2003 to February 2004, March 2011, and Oct 2014.

The change in Oct-2003 to Feb-2004 were most likely related to the fact that PPM EHR system were just started to be used in the LTHT. The change in March 2011 is when the PPM EHR system was started to be adopted to manage data in the whole hospital. In October 2014, PPM EHR system has been improved to join the Leeds Care Records (LCR) and connected to other providers such as General Practitioners (GPs), mental health, community and adult social care.

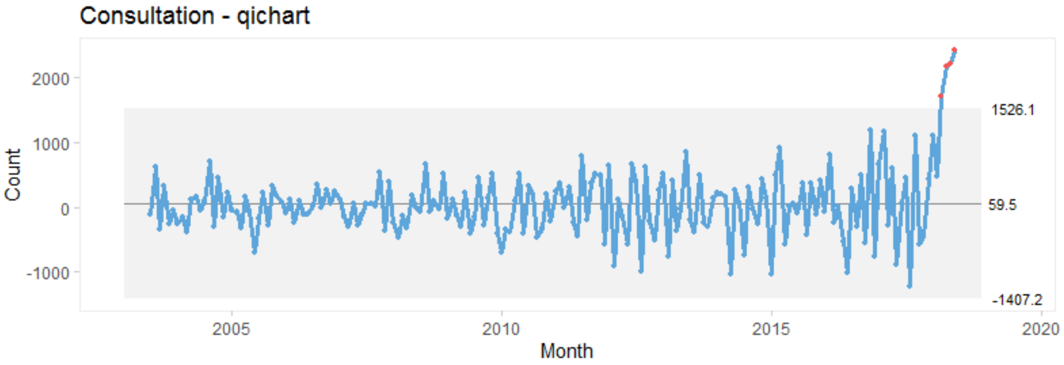
***Consultations** (Min 1; Med 5788; Mean 6459; Max 15025)*



Fitted trend: *(coef = 72.78571, adjusted* R^2 *= 0.7631, p-value < 2.2e-16)*



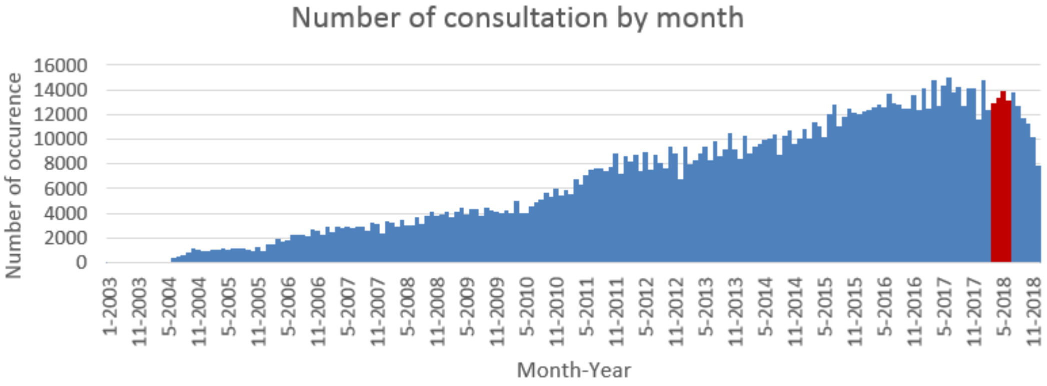
SPC to detect change points:



Change points:

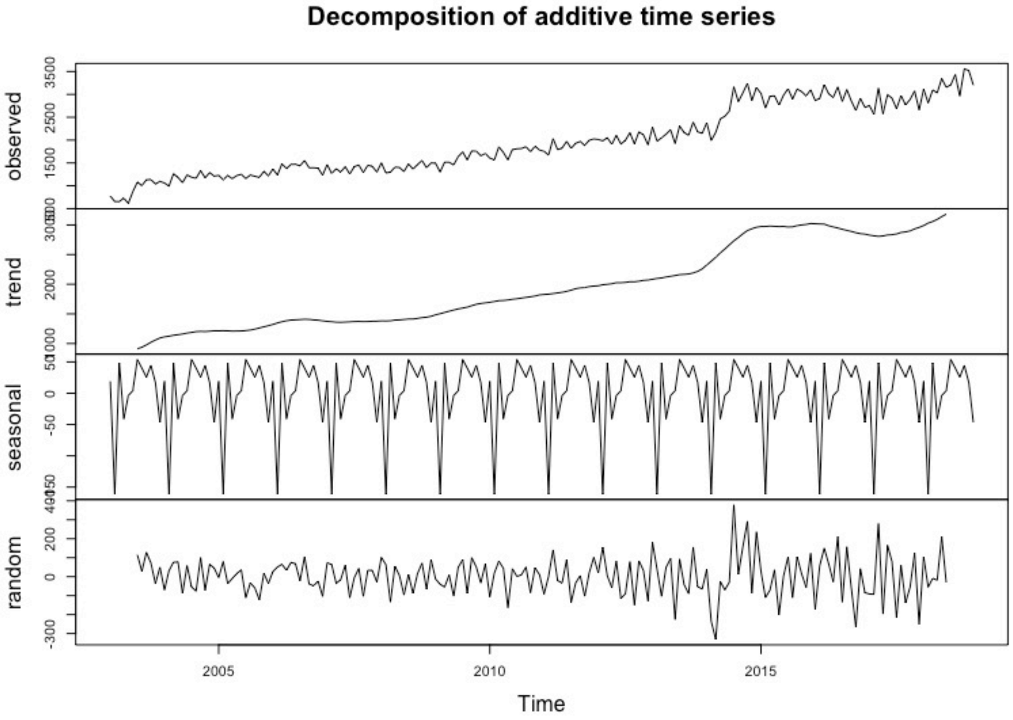
Month	3-2018	4-2018	5-2018	6-2018
value	1714.8	2182.6	2218.8	2418.6

Change points mapped into bar chart of monthly records:

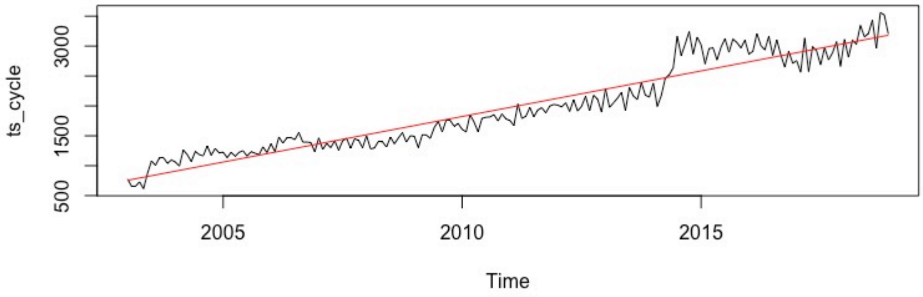


The trend of *Consultation* shows an increasing pattern by 72.78 records per month and is higher than the trend of *Diagnosis*. The change points detected in March–June 2018 were because *Consultation* was in the process of migration into a new system during 2018.

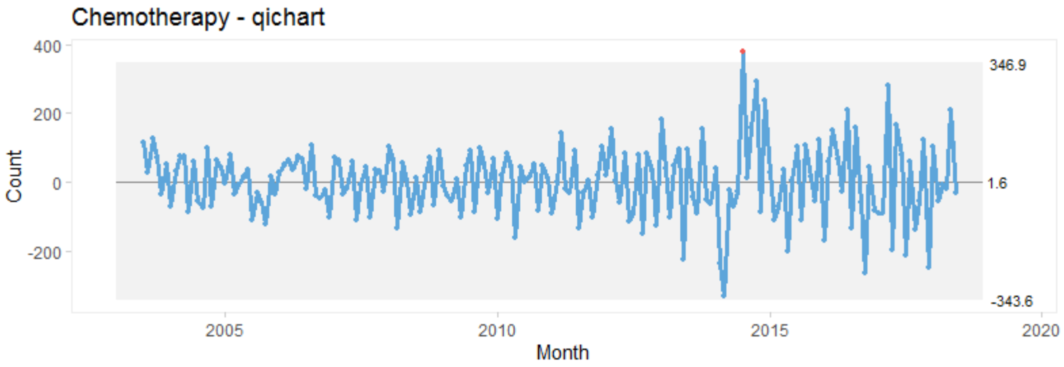
Chemotherapy *(Min 615; Med 1818; Mean 1971; Max 3559)*



Fitted trend: *(coef = 12.6943, adjusted R^2 = 0.9111, p-value < 2.2e-16)*



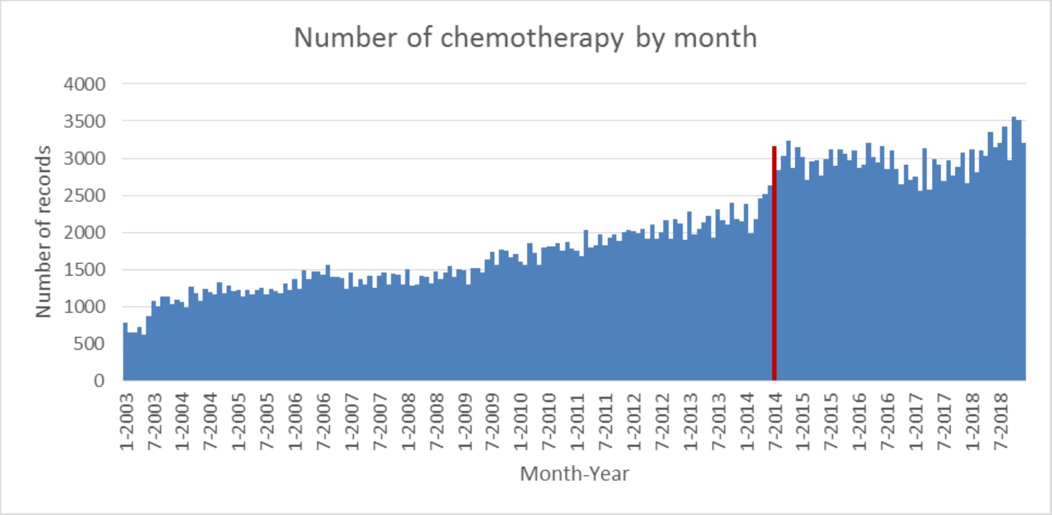
SPC to detect change points:



Change points:

Month	7-2014
value	378.5

Change points mapped into bar chart of monthly records:

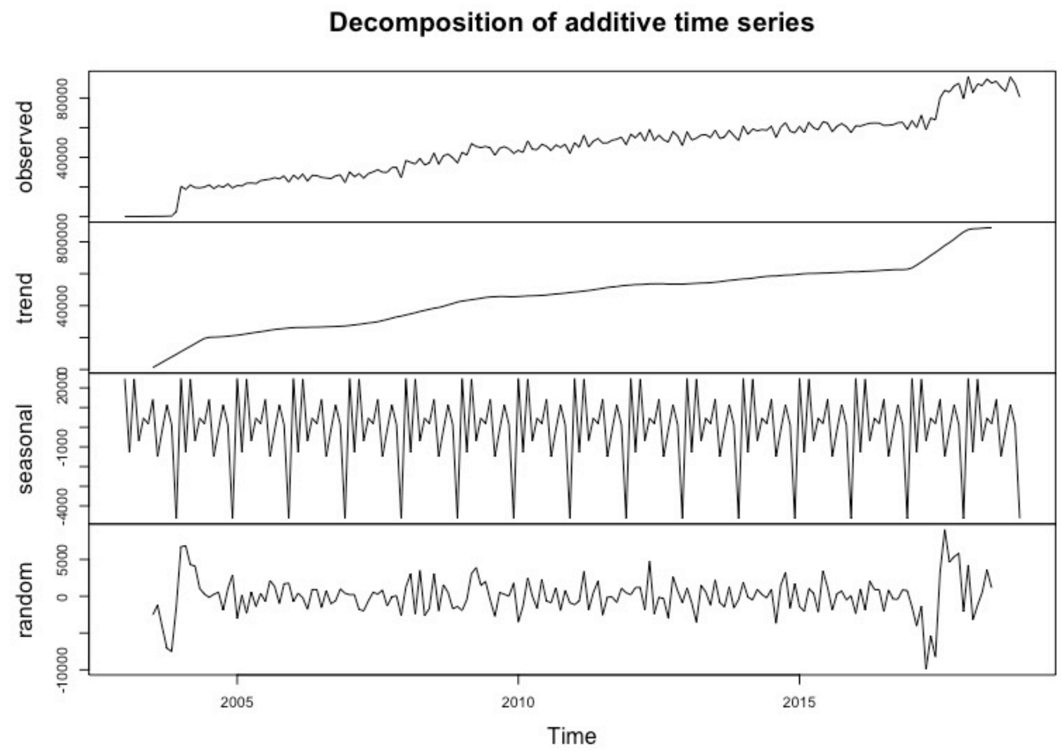


Chemotherapy is one of three cancer treatment types, along with surgery and radiotherapy. The trend of monthly records of chemotherapy in the PPM EHR system is increasing. One change point is detected in Jul-2014. This was discussed with clinical expert and was apparently because of a change in people authorized to do chemotherapy. Previously, it can only be done by clinicians in the gynaecology and gynaecology oncology departments. In 2014, clinicians in the haematology department was started to be authorized to do chemotherapy.

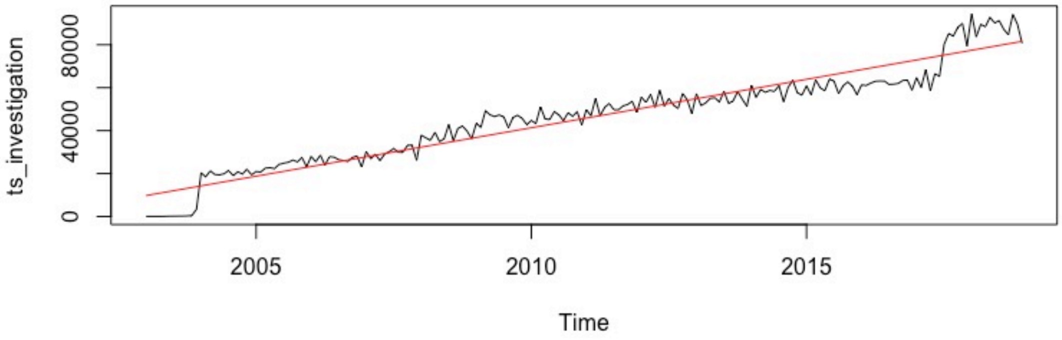
***Diagnosis** (Min 807; Med 3641; Mean 3512; Max 7568)*

Diagnosis can be said as the baseline in this experiment. This is because diagnosis was used in the selection criteria. More detailed results of the analysis of the monthly *Diagnosis* records are presented as an example in Section 6.6.3.

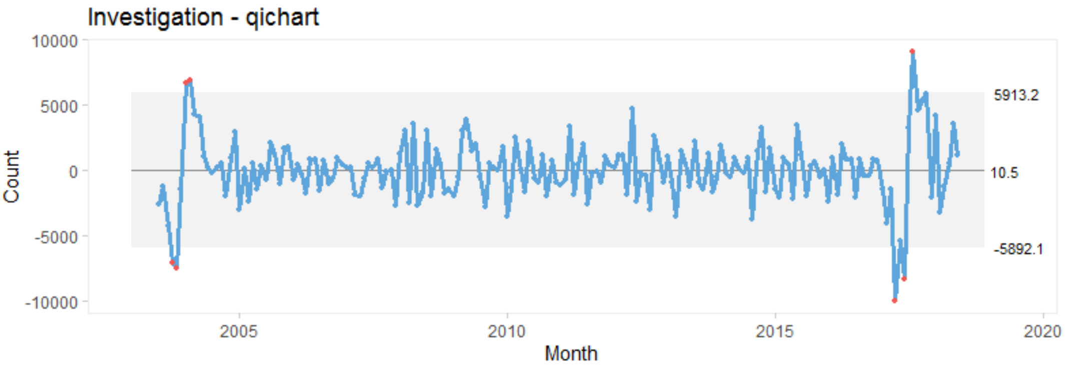
Investigation *(Min 38; Med 48662; Mean 45835; Max 94323)*



Fitted trend: *(coef = 376.0788, adjusted R^2 = 0.9156, p-value < 2.2e-16)*



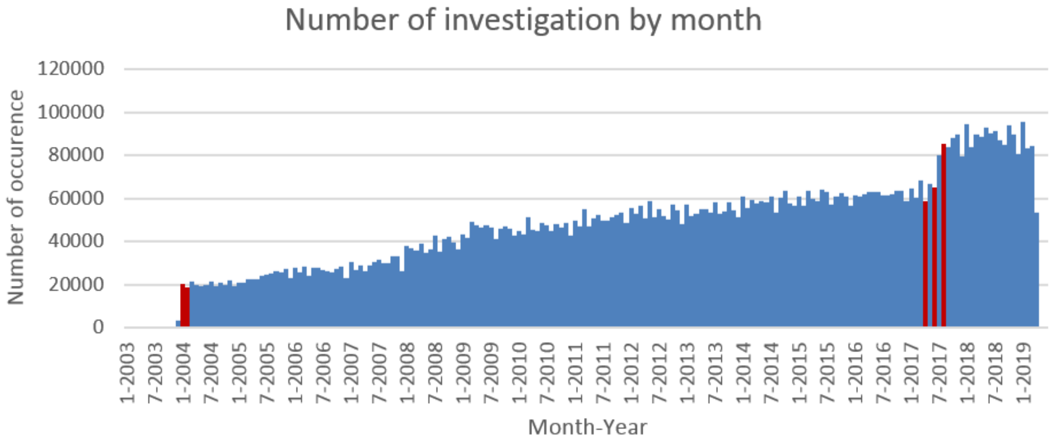
SPC to detect change points:



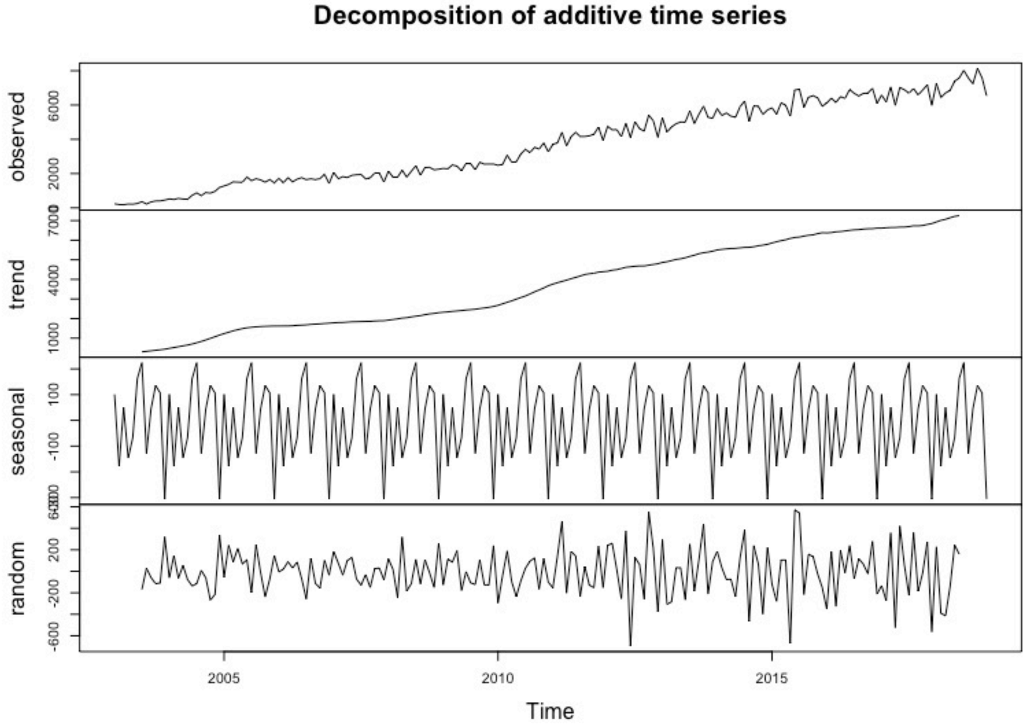
Change points:

Month	10-2003	11-2003	1-2004	2-2004	4-2017	6-2017	8-2017
value	-7072.8	-7504.5	6716.1	6852.2	-9932.0	-8249.7	9058.2

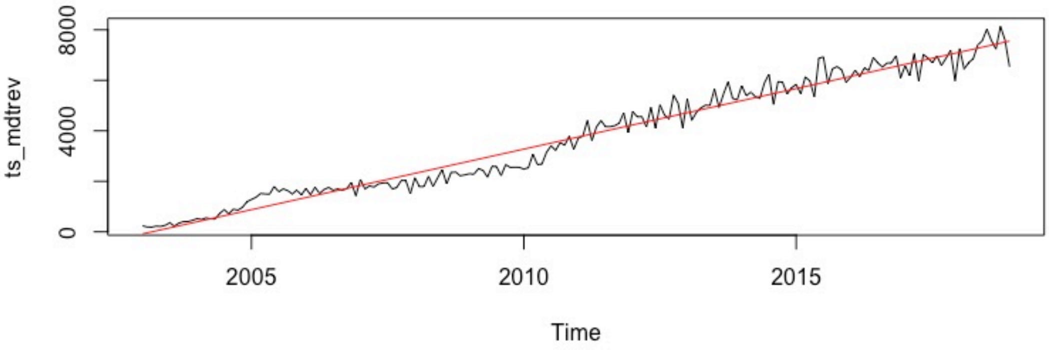
Change points mapped into bar chart of monthly records:



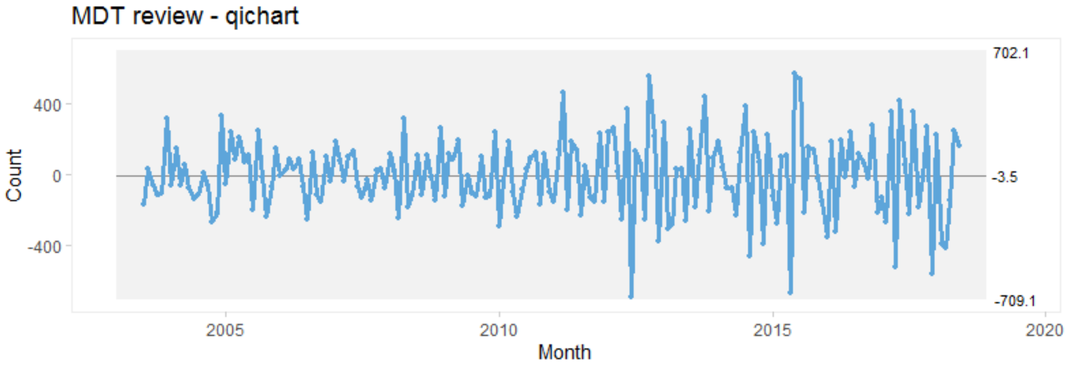
MDT review *(Min 183; Med 3658; Mean 3744; Max 8135)*



Fitted trend: *(coef = 39.9816, adjusted R^2 = 0.9678, p-value < 2.2e-16)*

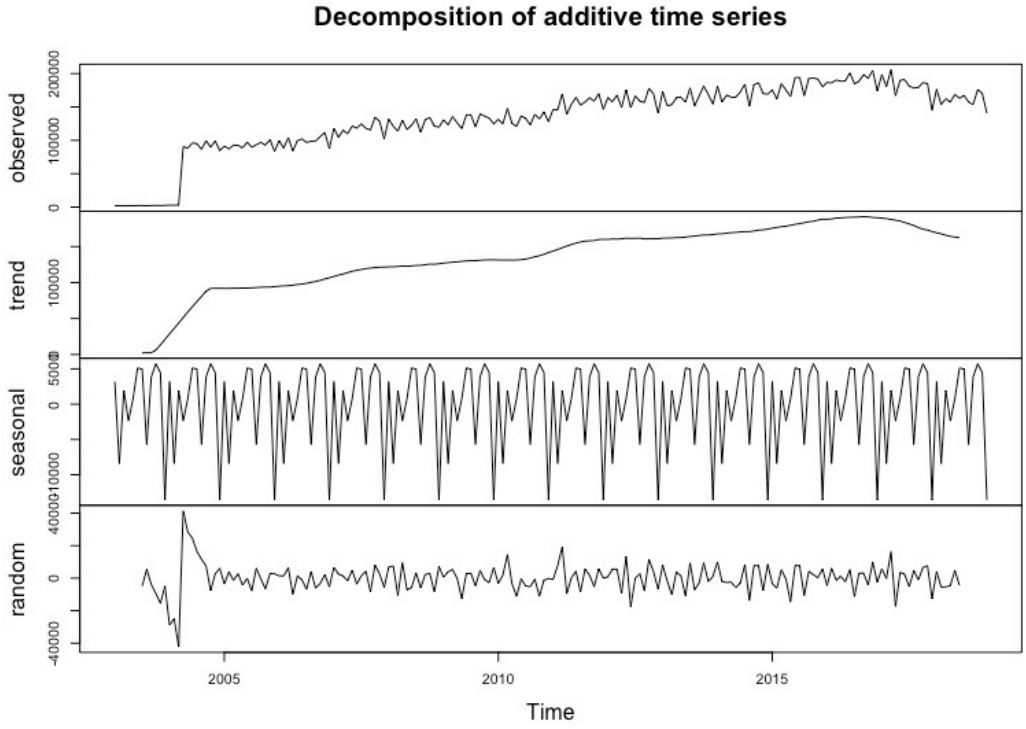


SPC to detect change points:

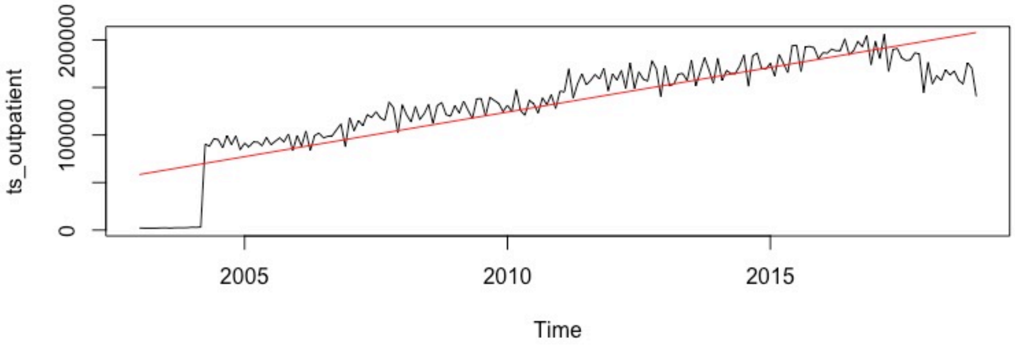


There is no change point detected based on the MDT review records. In the other word, the increasing trend is stable.

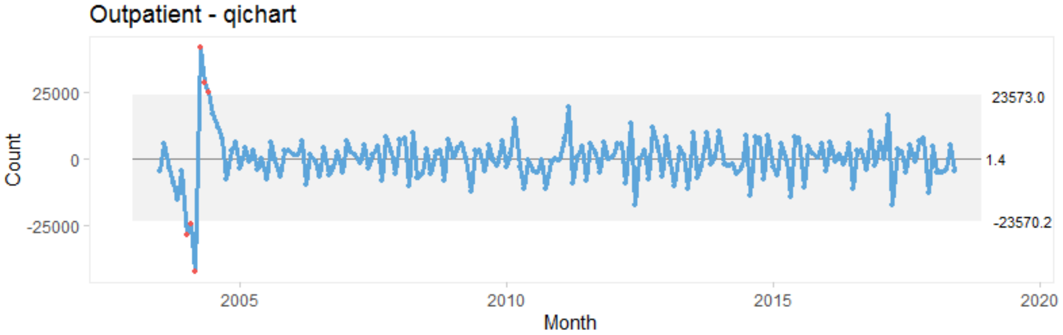
Outpatient *(Min 1977; Med 140584; Mean 133371; Max 205905)*



Fitted trend: *(coef = 791.9042, adjusted R^2 = 0.7555, p-value < 2.2e-16)*



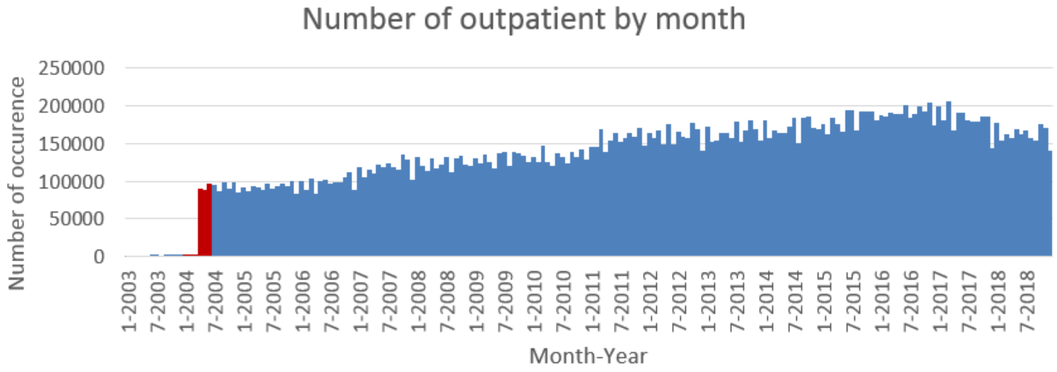
SPC to detect change points:



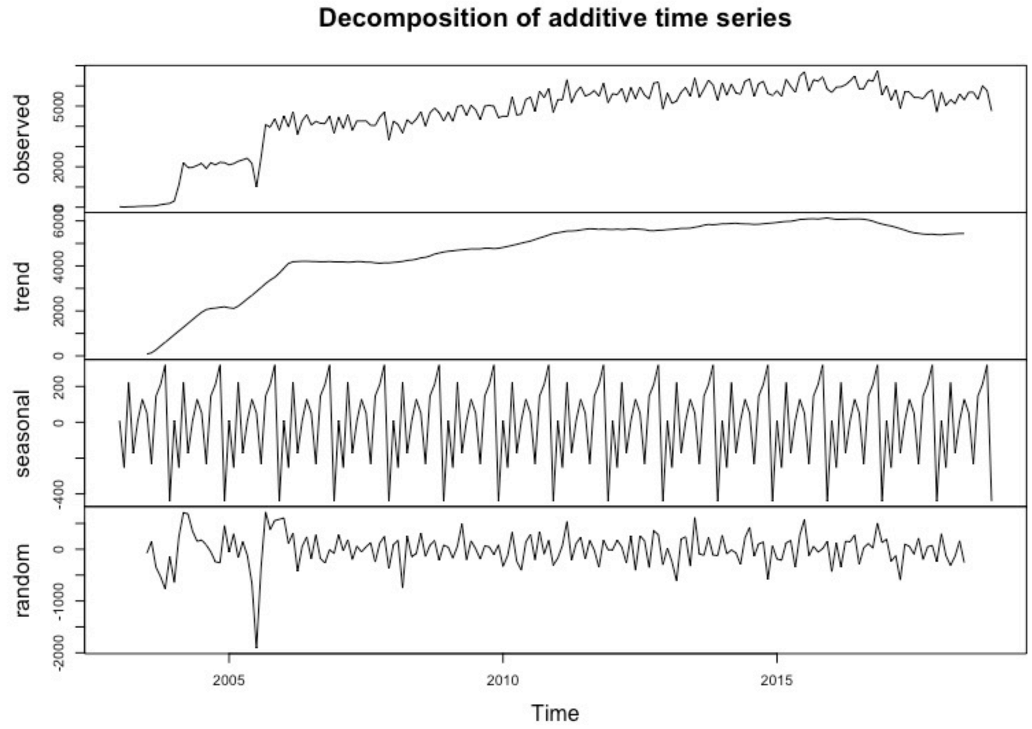
Change points:

Month	1-2004	2-2004	3-2004	4-2004	5-2004	6-2004
value	-28878,6	-24813	-42320.3	41464.8	28405.8	24579.7

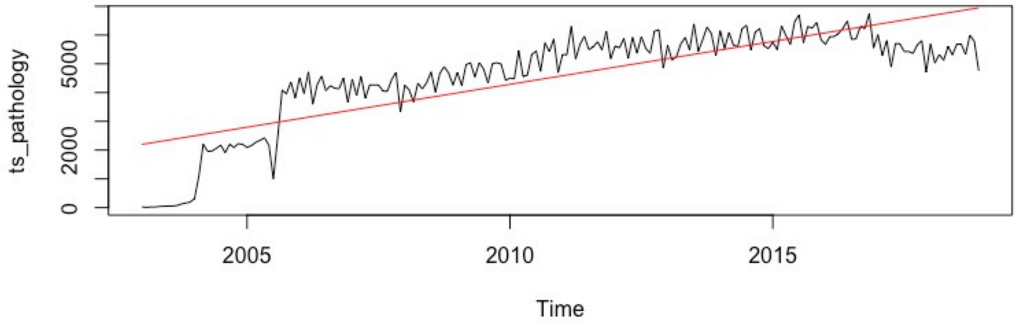
Change points mapped into bar chart of monthly records:



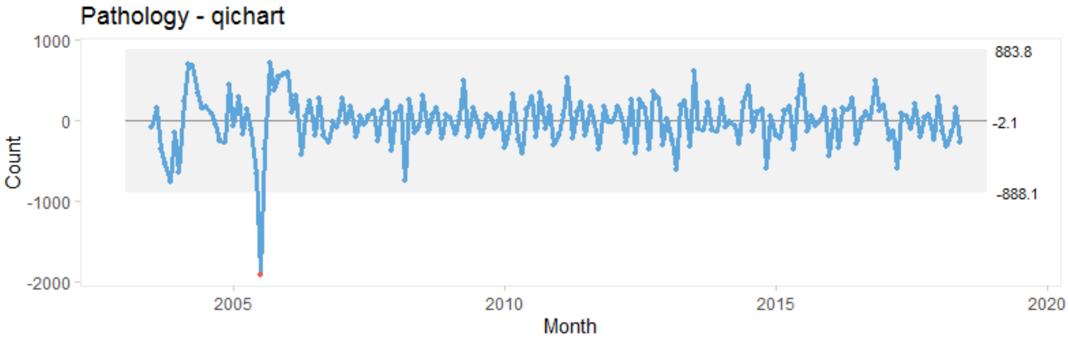
Pathology *(Min 8; Med 5126; Mean 4571; Max 6740)*



Fitted trend: *(coef = 24.8572, adjusted* R^2 *= 0.6626, p-value < 2.2e-16)*



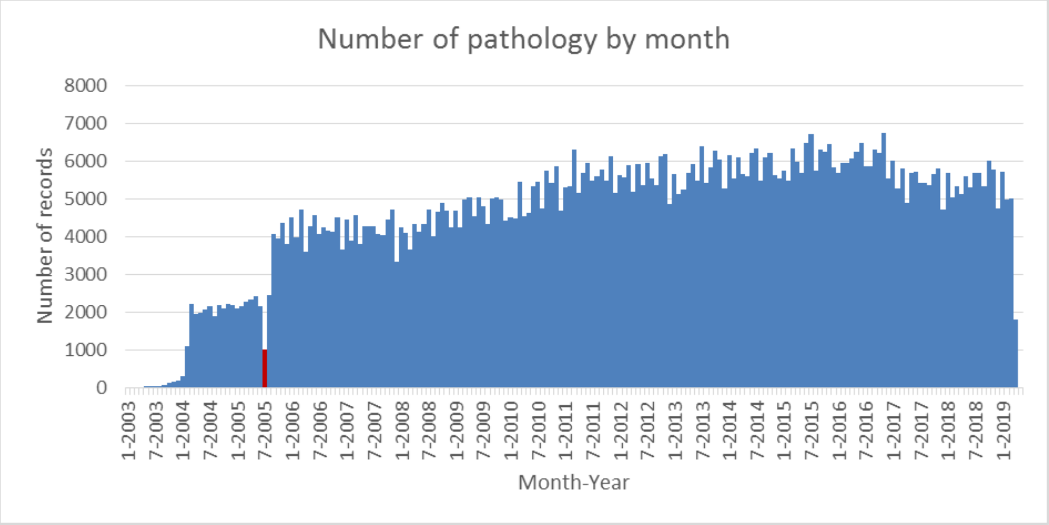
SPC to detect change points:



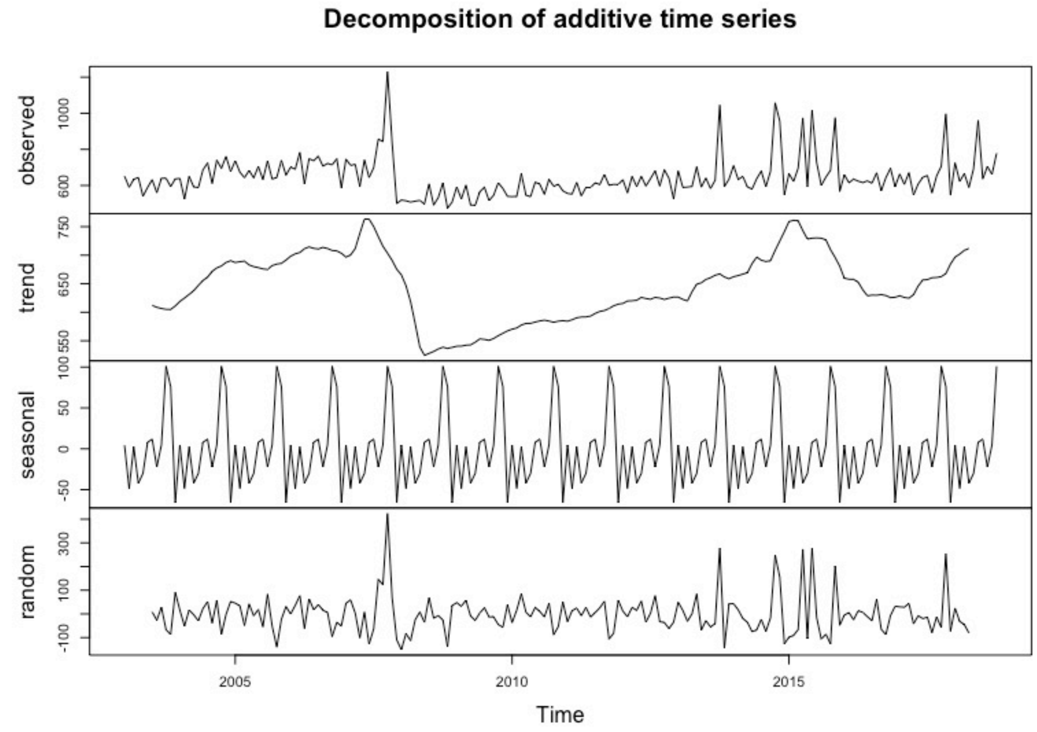
Change points:

Month	7-2005
value	-1911.8

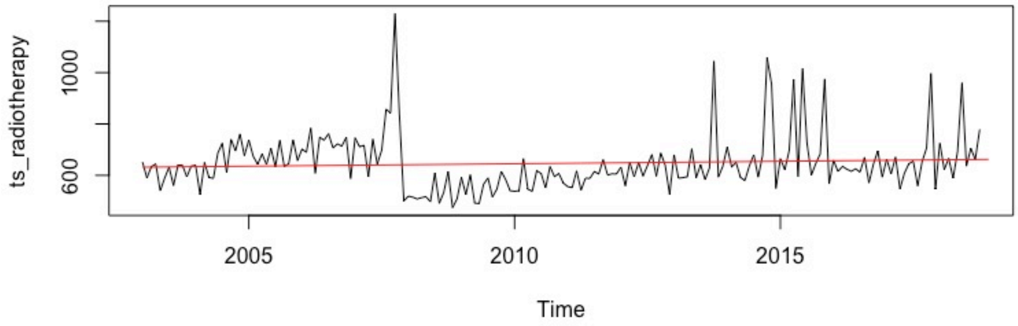
Change points mapped into bar chart of monthly records:



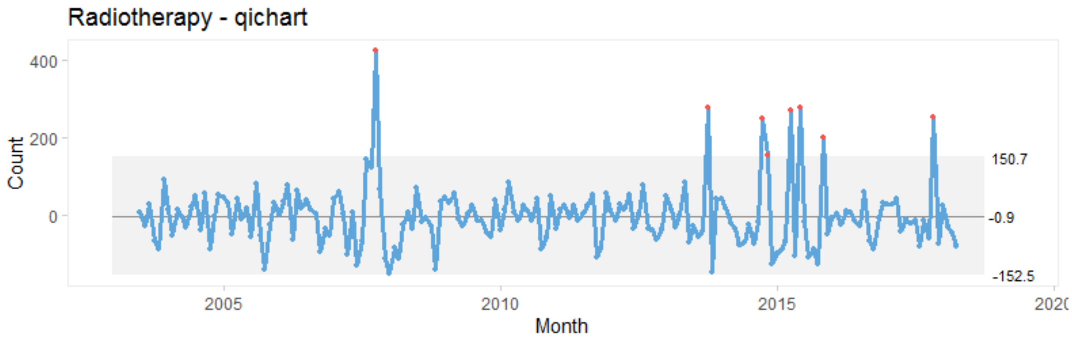
Radiotherapy *(Min 474; Med 633; Mean 647.1; Max 1229)*



Fitted trend: *(coef = 0.1599, adjusted R^2 = 0.00232, p-value = 0.2317)*



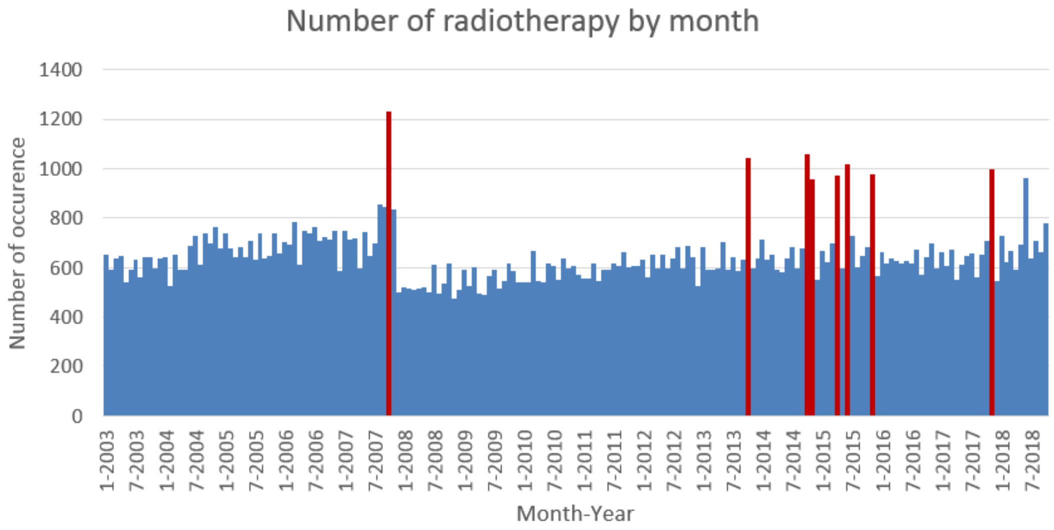
SPC to detect change points:



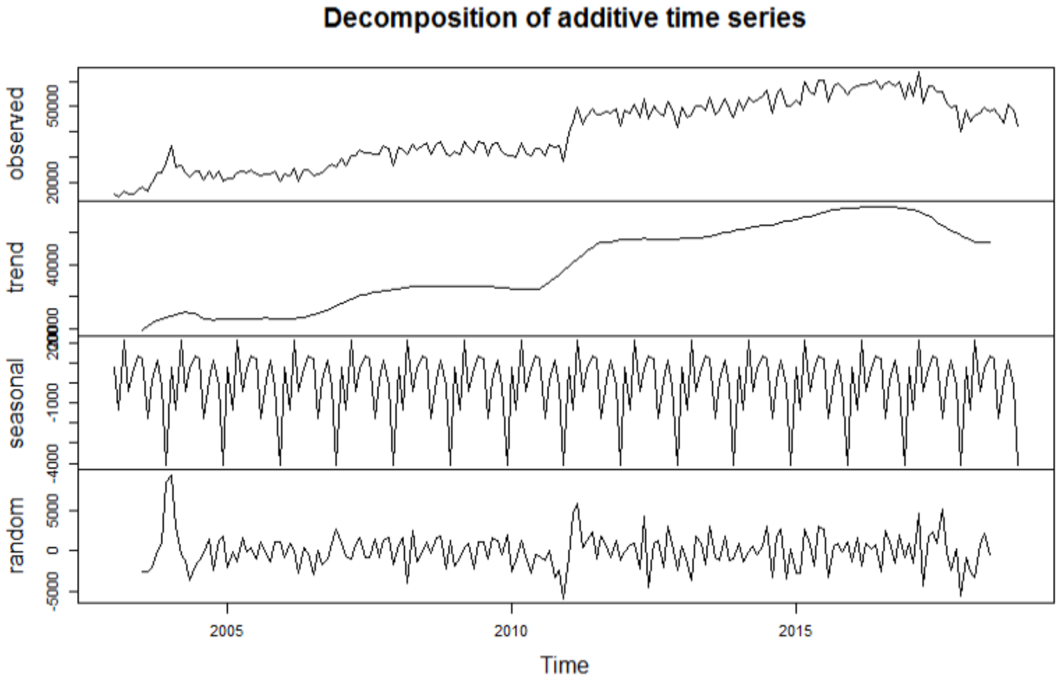
Change points:

Month	10-2007	10-2013	10-2014	11-2014	4-2015	6-2015	11-2015	11-2017
value	424.4	276.8	248.6	155.9	272.9	277.2	202.7	253.0

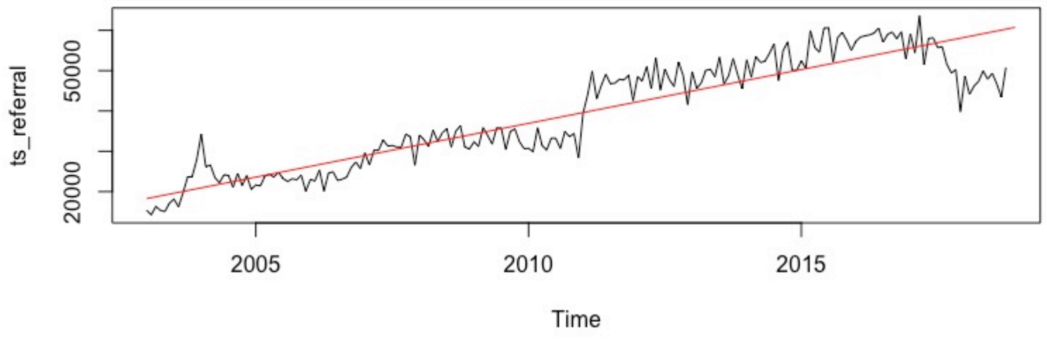
Change points mapped into bar chart of monthly records:



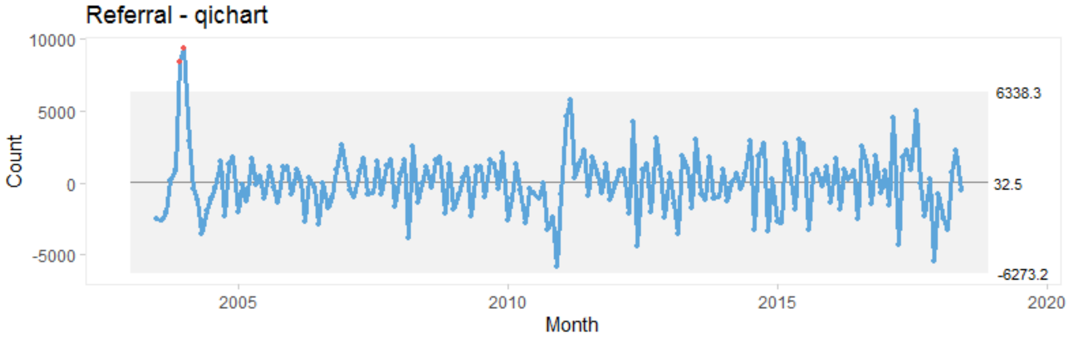
Referral *(Min 14223; Med 36088; Mean 39371; Max 63584)*



Fitted trend: *(coef = 222.2942, adjusted R^2 = 0.8401, p-value < 2.2e-16)*



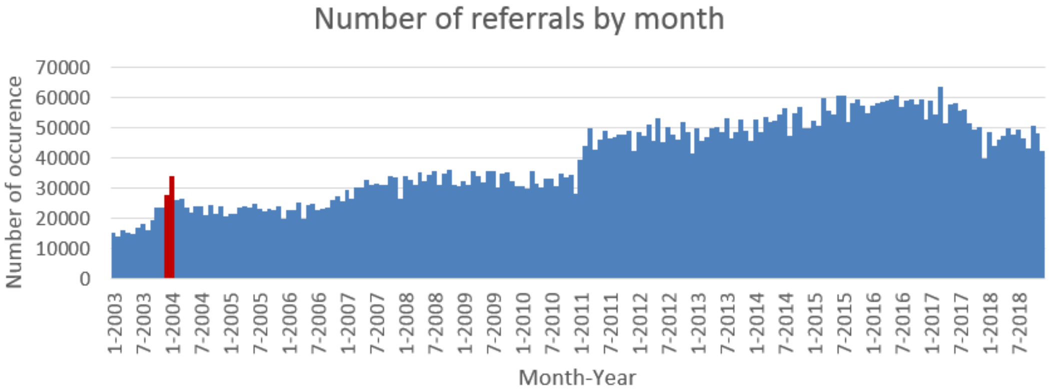
SPC to detect change points:



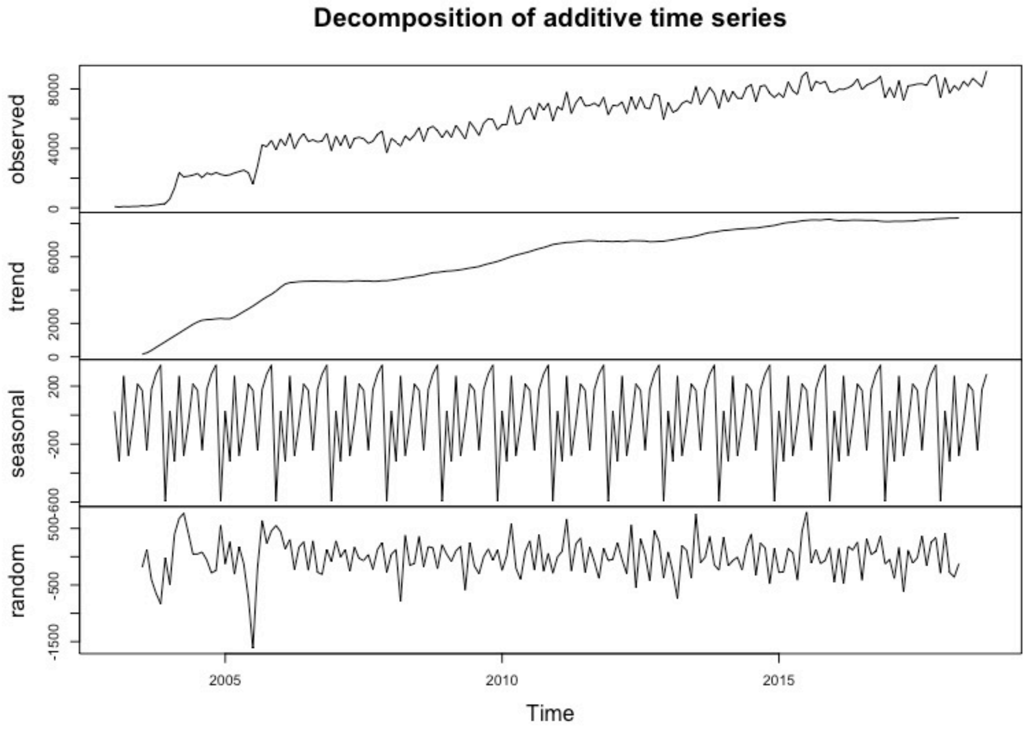
Change points:

Month	12-2003	1-2004
value	8397.9	9329.3

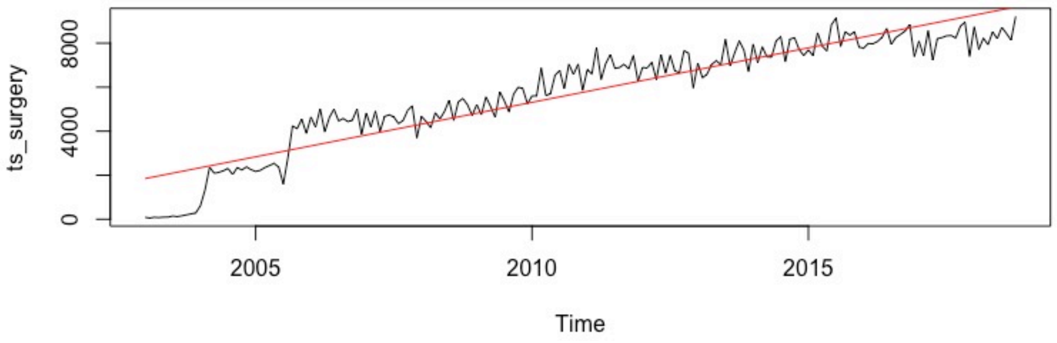
Change points mapped into bar chart of monthly records:



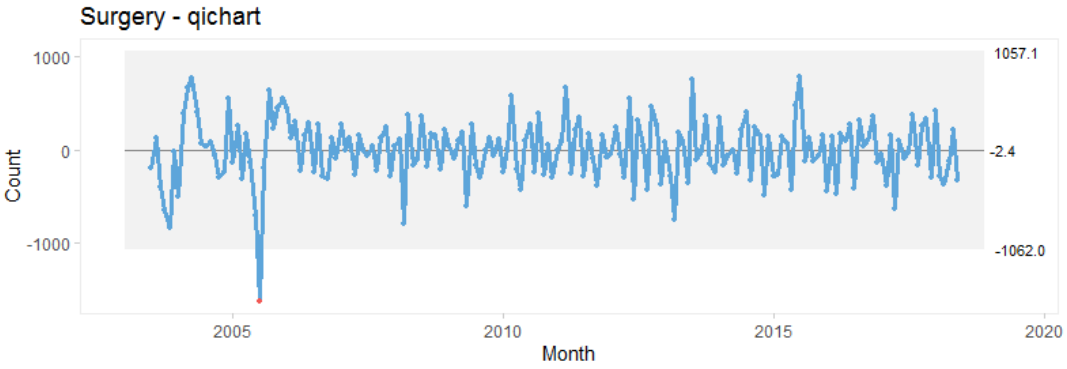
Surgery *(Min 56; Med 6480; Mean 5745; Max 9214)*



Fitted trend: *(coef = 41.24259, adjusted R^2 = 0.8681, p-value < 2.2e-16)*



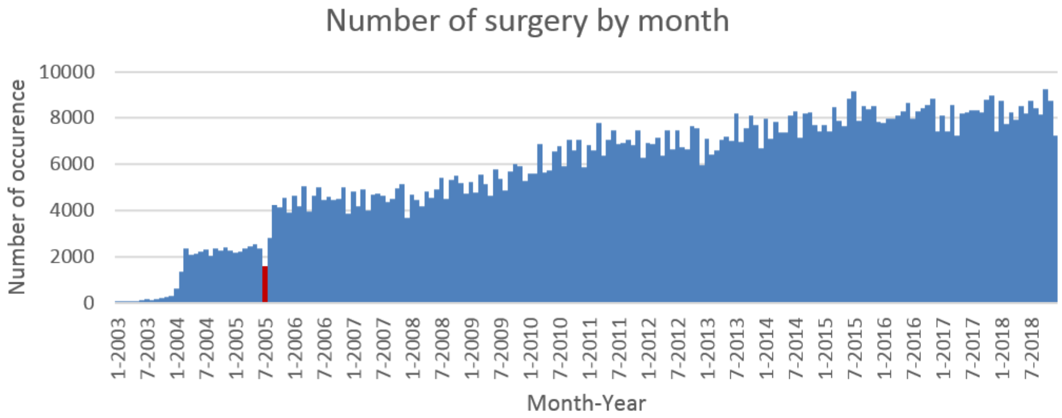
SPC to detect change points:



Change points:

Month	7-2005
value	-1616.724

Change points mapped into bar chart of monthly records:



Conclusion

Each activity has different pattern of trend, seasonal, and change point(s) detected based on this method. Further evaluation of the results have been presented in Section 6.6.4.

Appendix F
Overview of Plugins and Library

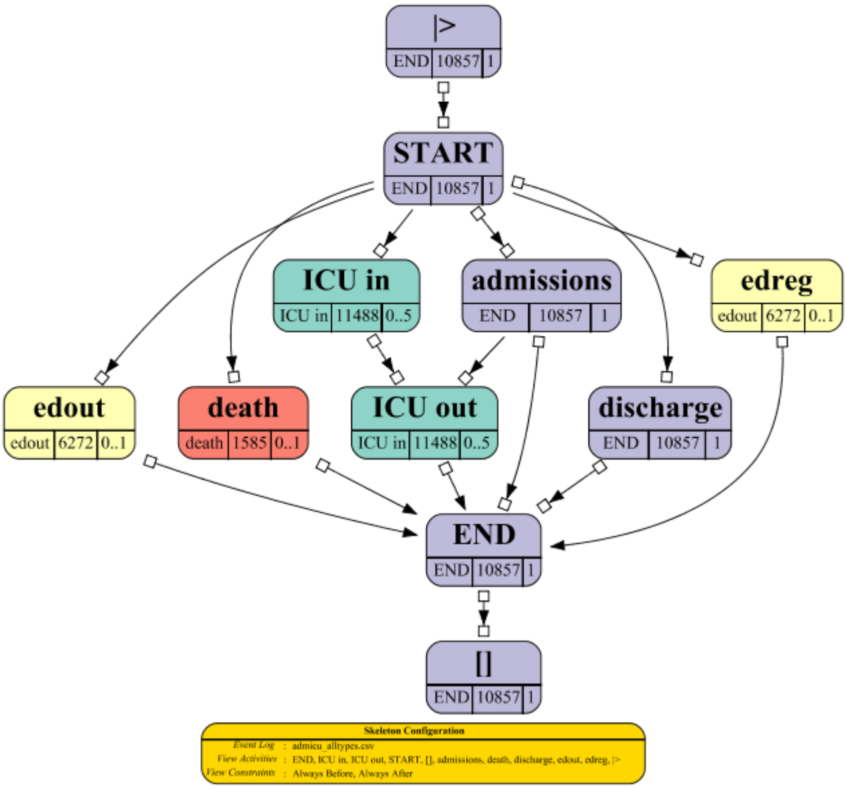
F.1 Log Skeletons: A Classification Approach to Process Discovery

Overview

This is a rule-based approach to visualize log skeleton model. This model is related to the Declare constraint model with a way to classify traces. The better the discovered model can classify trace conformance to the event log, the better the discovery algorithm is supposed to be.

Description

An event log can be seen as a bag (or multi-set) of sequences of activities (or activity traces). Every activity trace can be extended with an artificial start activity and an end activity (called as an extended log), to explicitly gives a better picture of the activity log in the end.



From the extended log, some relations can be defined, i.e. always after, always before, often next, often previous, never together, and next (one way or both ways). For example, the event log from experiment 1 in case study 1 as discussed in Section 4.3 can be visualised as the following log skeleton showing the 'always after' and 'always before' relations. Some important information visualised is that edout is not always after edreg, and discharge is not always after admissions. This indicates data quality issues of the completeness of edout and discharge.

Conclusion

This plugin is one of the analytics plugins in ProM. This plugin visualises relations of activities in the event log as a log skeleton.

F.2 bupaR

Overview

This is a package of functionalities for process analysis in R. This package is related to the other packages, including edeaR (exploratory and descriptive event-data analysis), processmapR, and processmonitR.

Description

The **bupaR** package is the core package of the framework that provides functions to create the objects and to support transformations such as mutate, filter, group_by, and mutate. The minimum required attributes to create an event log are case identifier, activity label, and timestamp. For example, the event log from experiment 2 in case study 3 as discussed in Section 6.4 can be created in bupaR with the following sintax.

-- *input: a dataframe* endcancer *of PatientID, Activity, Actor, Dated*

```
evlog <- endcancer %>%
      mutate(status=”complete”, activity_instance = 1:nrow(.))%>%
      mutate(Dated=as.Date(Dated, format=’%Y-%m-%d’))%>%
      eventlog(case_id=”PatientID”,activity_id=”Activity”,
      activity_instance_id = “activity_instance”,
      lifecycle_id = ”status”, timestamp = “Dated”,
      resource_id = “Actor”)
```

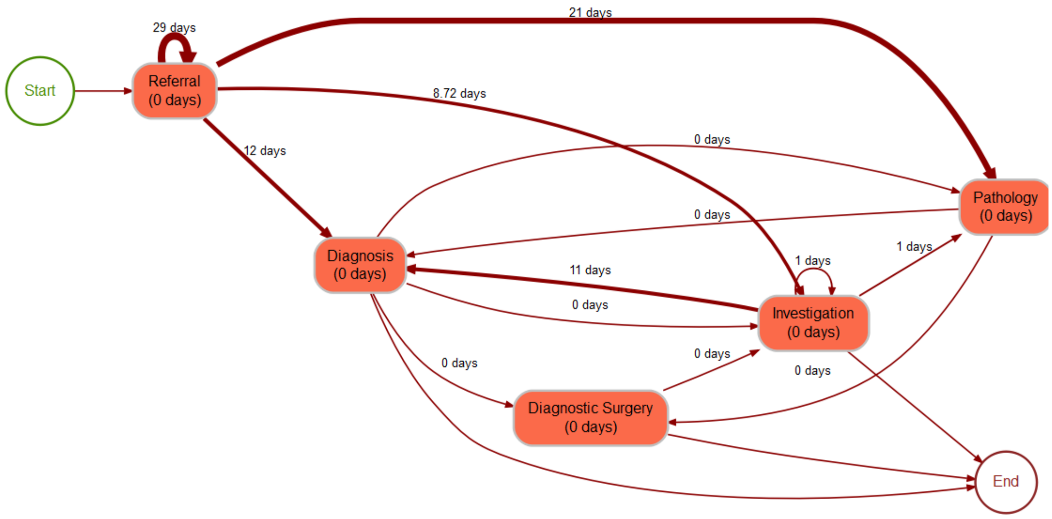
The **edeaR** package provides methods for describing and selecting process data, and for preparing event log data for process mining. Event data can be explored based on the time, organisational, and structuredness perspectives. Time perspective includes three metrics, i.e. throughput time, processing time, and idle time. Organisational perspective includes resource frequency, resource involvement, and resource specialisation metrics. For example, to analyse throughput time of processes in an event log, the following sintax can be used.

```
evlog %>% throughput_time (“activity”) %>% plot
```

The **processmapR** package provides several visualisations of the data, such as process maps and dotted chart. The visualisations are customisable to support different needs in the study. This includes options to present the absolute/ relative frequency or performance profile in term of mean, median, or custom profile in the process model. For example, the pathways of endometrial cancer from referral to diagnosis can be visualised using a process map presenting median durations. This can also be combined with a filtering function in **edeaR** to include only top 30% of the most frequent traces. The complete syntax is as follow.

```
evlog %>% filter_trace_frequency(percentage=0.3)
      %>% process_map(type=performance(median, units=”days”))
```

Output:



The **processmonitR** package provides a set of process dashboards. Some predefined dashboard are performance dashboard, activity dashboard, rework dashboard, and resource dashboard. Those dashboards combined some functions in the bupaR packages to provide more interactive visualisation of data and process of interest.

Conclusion

This package has been useful in the PPM Cancer case study in this thesis. Data description of event log has been made easy by bupaR with summary() sintax. The process discovery used in this package is visualised in state transition diagram. The analysis has been combined with many other basic packages in R such as stats, ggplot, and sqldf.