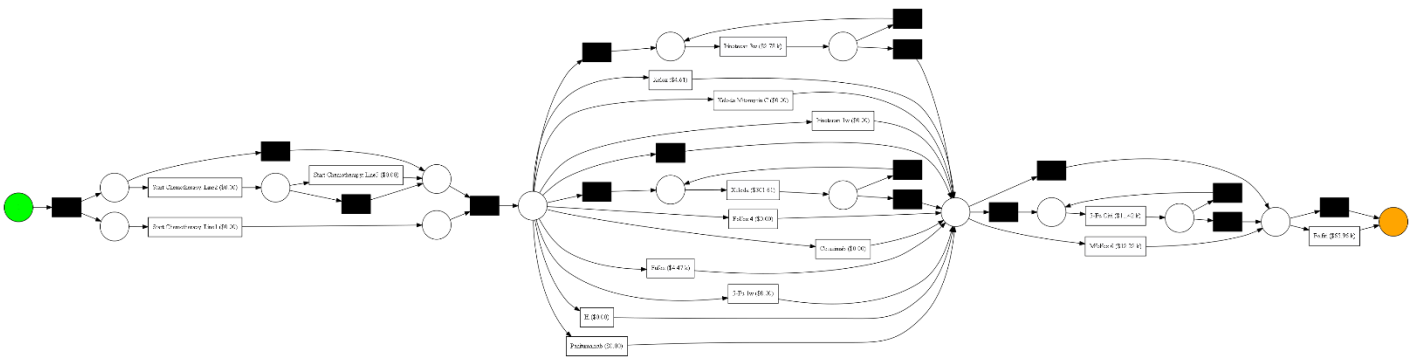
# ANALYSIS OF THE COSTS OF CARE FOR COLORECTAL CANCER PATIENTS USING AUSTRALIAN MULTI-CENTRE LINKED DATA

**Author: S.T. Relijveld**
**University of Twente**

**Cancer Health Services Research**
**Melbourne School of Population & Global Health**
**University of Melbourne**

**Master Thesis Industrial Engineering & Management, University of Twente**
**August 2021**



# ANALYSIS OF THE COSTS OF CARE FOR COLORECTAL CANCER PATIENTS USING AUSTRALIAN MULTI-CENTRE LINKED DATA

**6 August 2021**

**Author**
S.T Relijveld (Sven)
s1460986
University of Twente, Enschede, the Netherlands
Faculty of Behavioral Management & Social Sciences
Master program in Industrial Engineering & Management
Specialisation track in Health Care Technology & Management

**Graduation organisation**
Cancer Health Services Research Unit
Victorian Comprehensive Cancer Centre
Level 13, 305 Grattan Street, Melbourne, Victoria, Australia
The University of Melbourne
Faculty of Medicine, Dentistry and Health Sciences
Melbourne School of Population and Global Health

**University supervisors**
Prof. Dr. M.J. IJzerman (Maarten)
Faculty of Behavior Management and Social Sciences
Department of Health Technology and Services Research

Dr. D. Demirtas (Derya)
Faculty of Behavior Management and Social Sciences
Department Industrial Engineering and Business Information Systems

**External supervisors**
Dr. F. Franchini (Fanny)
Research Fellow
Cancer Health Services Research Unit
Melbourne School of Population and Global Health

Dr. M.Tew (Michelle)
Research Fellow
Health Economics Unit
Melbourne School of Population and Global Health

# MANAGEMENT SUMMARY

This proof-of-concept study demonstrates the use of Process Mining to analyse differences in care provided to colorectal cancer patients and the associated costs. The project was conducted within the Cancer Health Services Research Unit (HSRU), part of the University of Melbourne's Centre for Cancer Research (UMCCR) and the Melbourne School of Population & Global Health (MSPGH) and is part of a larger project on analyzing disparities in outcomes for colorectal cancer patients. In this study, real-world de-identified patient data was used from patients treated within three Australian hospital (groups) in the Melbourne metropolitan area and linked to primary care data.

To investigates differences between care and costs of care, a novel branch of data science named Process Mining is used, showing its value in automatic modelling of care pathways, quantitative support of model quality, and through insightful and partially interactive visualizations. Previous research into Process Mining in healthcare, focused mainly on single hospitals or groups with the same electronic health record system, while this research uses linked data and evaluates the entire pathway.

The main research question is formulated as:

*"How can Process Mining be applied to derive care pathways and analyse the costs of care provided to CRC patients in these care pathways?"*

To answer this question, a workflow for applying process mining is designed. This workflow includes selection of a certain population, applying an algorithm to derive process models (Discovery), check the quality of the models (Conformance) and extend the models to give better insights (Enhancement). To analyse the cost, a custom algorithm is designed, that sums the costs of each patient in each care activity, providing insights into where in the pathway certain costs are incurred.

The workflow was applied to a cohort of 7734 patients from the Peter MacCallum Cancer Centre (n=218), Western Health (n=4721) and the Royal Melbourne Hospital (n=2795). These patients are also included in the ACCORD clinical colorectal cancer registry, and subsequently linked to the Victorian Admitted Episodes Dataset (VAED) containing hospital records, the General Practitioner's primary care database Medicine Insight (NPS) and the registry Treatment of Recurrent and Advanced Colorectal Cancer (TRACC). This selection resulted in a final population of 4246 patients.

The care patients received was transcribed into an event log. Which is a data storage format, suitable for process mining. For the hospital care, names from the hospital's Diagnosis Related Groups (DRG's) was used, for the primary care from descriptions in the Medicare Benefits Schedule (MBS) and for the medication by the registered name in the Pharmaceutical Benefit Scheme (PBS). The costs for the hospital's care was calculated with the Weighted Inlier Equivalent Separation (WIES) scheme, while costs for primary care and medication were based on the prices in MBS and PBS respectively.

The resulting pathways and quality metrics are displayed in an interactive online app. A comparison is made for all phases and a case study is performed to find the differences between care and costs of care between colon cancer patients in different stages of their disease. In this case study, we found that hospital admissions are the costliest aspect for all stages, and that most disparities between the stages occurs in chemotherapy, where there is a large difference between costs of chemotherapy regimen MFOLFOX 6. Based on the results it can be concluded that the designed workflow including the Process Mining techniques can aid health services researchers in analysing differences in care provision and costs between groups of patients.

Within this study, Process Mining proved to be a value-adding method for providing insights on differences between patient groups in complex care. This methodology is more data-driven, contrary to consensus-based guidelines like Optimal Care Pathways, and displays actual provided care on a detailed level, including deviations that doctors routinely make to accommodate for patient characteristics and -preferences. The field of Process Mining is expected to grow rapidly over the next years and to be applied in case studies in health services research and other domains within the healthcare sector.

Additional research should focus on the primary care, as in this study, the number of patients linked to the primary care dataset was relatively low. The absolute number of patients linked to the primary care dataset (1106) was relatively low. Also, this number became even lower when these patients were eventually included based on their symptoms (187 compared to the 3233 in hospital care), which could reduce the validity of the obtained models. Furthermore, the process models for primary care obtained in this study could be improved by implementing a better suited classification scheme or by implementing a Natural Language Processing component in the workflow, to cluster groups of activities that are relatively the same together under a single name. This would yield better interpretable models, as well more valid models that describe the actual provided care.

# ACKNOWLEDGEMENTS

I would like to thank those who provided me with their expertise, and guidance that made it possible to complete this master's thesis project.

Dr. Fanny Franchini

Prof. Dr. Maarten IJzerman

Dr. Michelle Tew

Dr. Derya Demirtas

Allison Drosdowsky

Dr. Chris Kearney

Dr. Hui-Li Wong

Dr. Peter Gibbs

Staff of the Cancer Health Services Research Unit

Staff of the BioGrid Australia Team

# CONTENTS

# LIST OF TABLES

# LIST OF FIGURES

# LIST OF ABBREVIATIONS AND ACRONYMS

| ***Abbreviation*** | ***Definition*** |
|---|---|
| (m)CRC | (Metastatic) Colorectal Cancer |
| ACCORD | Australian Comprehensive Cancer Outcomes And Research Database |
| ACPS | Australian Clinico-Pathological Staging System |
| AJCC | American Joint Committee On Cancer |
| ASR | Age-Standardized Rate |
| CEA | Carcinoembryonic Antigen |
| CPI | Consumer Price Index |
| CT | Computed Tomography |
| DFG | Direct-Follow Graph |
| DPN | Data-aware Petri-Nets |
| DRG | Diagnosis Related Group |
| EGFR | Epidermal Growth Factor Receptor |
| EHR | Electronic Health Records |
| EPC | Event-Driven Process Chains |
| ERP | Enterprise Resource Planning |
| GIST | Gastrointestinal Stromal Tumours |
| HSRU | (Cancer) Health Services Research Unit |
| IEEE | Institute Of Electrical And Electronics Engineers |
| iFOBt | Immunochemical Faecal Occult Blood Test |
| IHPA | Independent Healthcare Pricing Authority |
| MBS | Medicare Benefits Schedule |
| MRI | Magnetic Resonance Imaging |
| MSPGH | Melbourne School of Population & Global Health |
| NBCSP | National Bowel Cancer Screening Program |
| NEP | National Efficient Price |
| NPS | National Prescribing Service |
| NWAU | National Weighted Activity Unit |
| PBS | Pharmaceutical Benefit Scheme |
| PET | Positron Emission Tomography |
| PICC | Peripherally Inserted Central Catheter |
| SES | Socio-Economic Status |
| SSP | Sessile Serrated Polyps |
| TNM | Tumor-Node-Metastasis |
| TRACC | Treatment Of Recurrent And Advanced Colorectal Cancer |
| TSA | Traditional Serrated Adenomas |
| UICC | Union Internationale Contre Le Cancer |
| UMCCR | University Of Melbourne's Centre For Cancer Research |
| USI | Unique Swap Identifier |
| VAED | Victorian Admitted Episodes Dataset |
| VEGF | Endothelial Growth Factor |
| VEMD | Victorian Emergency Minimum Dataset |
| VINAH | Victoria Integrated Non-Admitted Health |
| WIES | Weighted Inlier Equivalent Separation |
| YAWL | Yet Another Workflow Language |

# List of Appendices

# 0 PREFACE

This thesis is written as a graduation project for the master Industrial Engineering & Management, specialisation Healthcare Technology & Management, aiming to demonstrate the capabilities of Process Mining, a novel Data Science branch in the context of health services research. Specifically, it is used in the cost-analysis of care pathways for Australian colorectal cancer patients. The project was under the supervision of the Cancer Health Services Research Unit, part of the University of Melbourne's Centre for Cancer Research (UMCCR) and the Melbourne School of Population & Global Health (MSPGH).

The HSRU obtained ethical approval to use real-world deidentified patient data, collected by BioGrid, an Australian connectivity platform for medical data. In the thesis, we aim to apply process mining, a novel data mining subfield, suited to derive process models from low-level data to the continuum of care for patients with colorectal cancer (CRC), treated within the health centres of three Australian hospital(groups) in the Melbourne metropolitan area. Next, we use our established model to analyse and compare the (efficiency of) care and associated costs provided to CRC patients. For this project, we use data from the ACCORD colorectal cancer registry, the Victorian Admitted Episodes Dataset (VAED), the General Practitioner's primary care database Medicine Insight (NPS) and the registry Treatment of Recurrent and Advanced Colorectal Cancer (TRACC). All data used is related to a cohort of patients from the Peter MacCallum Cancer Centre (n=218), Western Health (n=4721) and the Royal Melbourne Hospital (n=2795), located in Victoria, Australia.

The thesis consists of eight chapters: in the first chapter, the disease context for colorectal cancer is introduced, as well as the motivation for this study and the objective of this study. This chapter is followed by a chapter giving an introduction in the field of Process Mining, concluding with a research question. The third chapter describes the methodology used to apply process mining in the context of deriving care pathways. Next, an algorithm for computing costs as an attribute for a care activity and concludes with a workflow to pre-process and to mine the consequent data to obtain pathways. The fourth chapter provides an experimental setup for analysing costs in the pathway of colorectal cancers with the linked data from BioGrid and allows comparison across different cohorts. In the fifth chapter, the resulting outcomes of the experimental setup are discussed to illustrate its value. The sixth chapter demonstrates the applicability of the methodology in a case study comparing cost and care differences in Colon cancer between cohorts with different ACPS stages. The seventh chapter answers the research question and discusses the results as well as the relevance of applying process mining in health care and future directions, and the eighth chapter concludes this thesis.

I have worked on this thesis from 2020-2021, a year disrupted by the coronavirus, leading to working remotely from the Netherlands. This was challenging at times, but I have persevered and with the conclusion of it, I will complete my Masters in Industrial Engineering & Management. I cannot forget to express my gratitude to my all supervisors for their help, insights and advice in all parts of my research. As well thanks to all my friends within 'Magnus', who have been a rocksteady support in a time, where the world seemed to be on fire constantly. A special thanks to Eiko Westerbeek, for reading through my entire thesis, giving valuable feedback on my wording and writing style. As well love and thanks for my family that have supported me throughout this journey and provided me with a roof over my head when I was unable to fly to Melbourne. And last but not least, my girlfriend Thirza, on who I've leaned on the most, giving me the motivation to keep pushing on and alleviate the stress.

I hope you enjoy reading my thesis!

6th of August 2021
Sven Relijveld

# 1 INTRODUCTION

This chapter first describes the context of the research conducted in this graduation project and provides an overview of the pathology, epidemiology, and clinical management of colorectal cancer. It is followed by a section on the motivation of using automated derivation of care pathways and an analysis on the costs of care provided throughout the care pathway and concludes with an objective for this study.

## 1.1 Research field and organisation

Health services research is the multidisciplinary field of scientific investigation that studies how social factors, financing systems, organizational structures and processes, health technologies and personal behaviors affect access to health care, the quality and cost of health care and ultimately our health and well-being [1]. The Cancer Health Services Research Unit (CHSRU), part of the University of Melbourne's Centre for Cancer Research (UMCCR) and the Melbourne School of Population & Global Health (MSPGH), has a specific focus on analysing the complex systems that revolve around cancer care.

All around the globe, incidence rates of cancer are rising and the costs associated for its treatment with it. To make the best use of the limited resources available, thorough analysis on the quality, outcomes and costs of care provided to cancer patients is vital. As cancer is a long-lasting or often chronic disease, the care associated with it, is not limited to a single healthcare organization and health services. Therefore, economics researchers focus more and more on all care provided to patients for specific disease type and its associated health problems: the (integrated) care pathway.

One of the current projects of the CHSRU is about analysing disparities in survival outcomes and health care resource utilisation in colorectal cancer. Colorectal cancer (CRC) is a cancer of large intestine, comprising the bowel, sigmoid and the rectum. It is the world's fourth most commonly diagnosed cancer and the third most deadly [2]. Colorectal cancer incidence is rising worldwide, linked to increased carcinogenic risk factors such as obesity, sedentary lifestyle, red meat consumption, alcohol, and tobacco. Mortality has been decreasing in developed nations due to advances in early detection with population screening and improved treatment options [3].

This thesis aims to extend on the colorectal cancer project of the Cancer Health Services Research Unit, by showcasing the utility of a novel field of data science, called Process Mining, which can derive process models as an abstraction from raw tabular data. Process Mining techniques may improve on the insights into the care provided in complex care pathways such as the care provided to patients with colorectal cancer. The following paragraphs provide context on the disease of colorectal cancer, its epidemiology, and the current standard of care according to the Australian guidelines.

## 1.2 Pathology

CRC is the collective name of the cancerous growths in either a part of the large intestine or in the rectum [4], [5]. Most colorectal cancers start out as growths or *polyps* on the inner lining or epithelial cells of the colon or rectum and are often qualified as pre-cancerous growth [6], see Figure 1. The chance of a polyp turning into lower staged cancer depends on the type of polyp it is. The main groups of polyps are:

- **Adenomatous polyps (adenomas):** These types of polyps are made up of tissue that looks like the normal lining of the colon and might develop into cancerous tissue. Because of this, adenomas are called a pre-cancerous condition. The 3 types of adenomas, based on their growth pattern are tubular, villous, and tubulovillous.
- **Hyperplastic polyps and inflammatory polyps:** These polyps are more common, but in general they are not pre-cancerous. Some people with large (more than 1cm) hyperplastic polyps might need colorectal cancer screening with colonoscopy more often.
- **Sessile serrated polyps (SSP) and traditional serrated adenomas (TSA):** These polyps are often treated like adenomas because they have a higher risk of colorectal cancer.

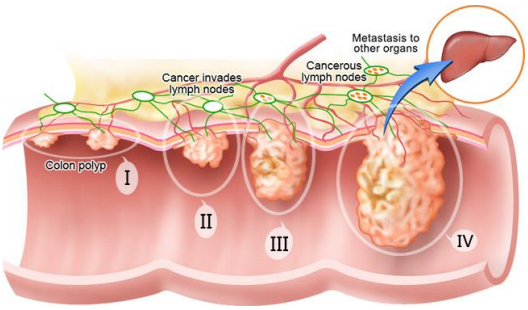



***Figure 1: Development of colorectal cancer from polyp in the inner lining of the colon.***

Over time, a polyp may acquire founder mutations that will promote further proliferation and allow the transitioning to a cancerous growth. Over time, it can grow into the wall of the colon or rectum, which consists of many layers. Colorectal cancer starts in the innermost layer (the mucosa) and can grow outward through some or all the other layers. In advanced CRC, cancer may spread to other sites of the body through the lymphatic or vascular system, where it forms metastases. The five-year survival rate for patients is greatly reduced if metastases are found. Depending on the cells from which the cancer originates, four major subtypes of CRC are identified. Adenocarcinomas grow from the epithelial cells, gastrointestinal carcinoid tumours from hormone-producing cells, gastrointestinal stromal tumours (GISTs) from interstitial cells, lymphomas from lymph-producing cells and sarcomas from blood vessels, muscle layers or connective tissue. Adenocarcinomas have by far the highest prevalence, with 95% of all CRC-cases being adenocarcinoma [7]and treatment options for the other types can differ from treating adenocarcinomas. The prognosis for patients with carcinomas of the colon is poor (5 years relative survival < 30%), better for lymphomas and sarcomas, and best for carcinoid tumours. There has been no significant change over time in long-term survival rates for any of the 4 histological subtypes [8].

## 1.3 Epidemiology

Worldwide, the incidence of colorectal cancer is estimated to be 1,096,601 in 2018, with approximately 881,000 deaths [2]. The highest colorectal cancer incidence rates (see Figure 2) are found in parts of Europe, Australia/New Zealand, Northern America, and Eastern Asia, while fairly low rates of both colon and rectal cancer are found in most regions of Africa and in Southern Asia. In Australia, the estimated number of new cases is approximately 15,500 in 2020, making it the fourth most common cancer and with 5,322 deaths it is the second most lethal cancer, after lung cancer [9].

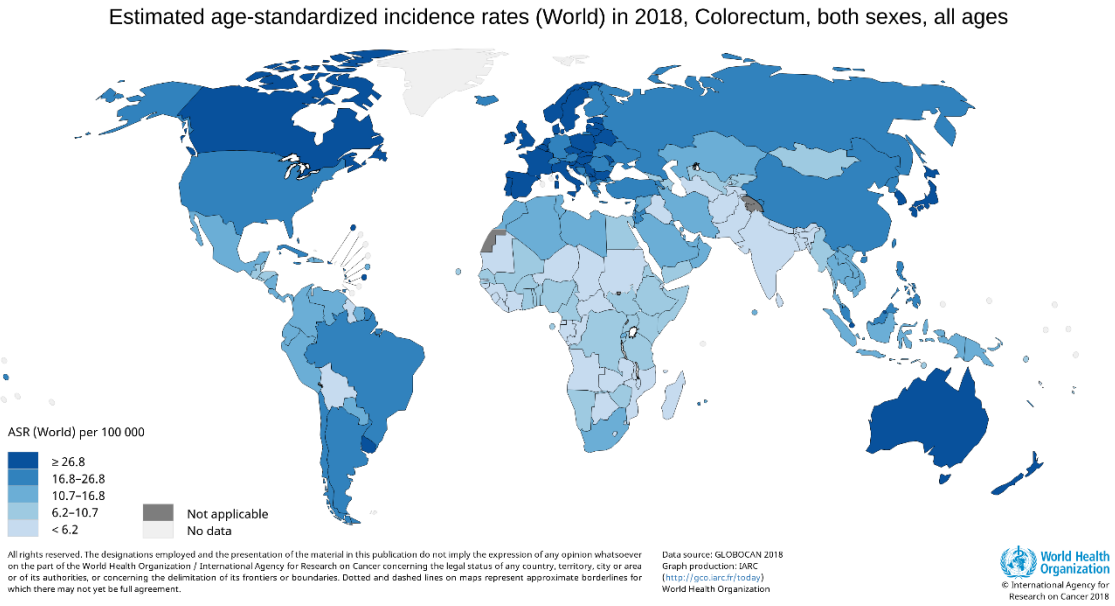



***Figure 2: World incidence rates of colorectal cancers: the age-standardized rate (ASR) per 100,000 inhabitants is shown in increasing intensity of blue. Source: Global Cancer Observatory***

CRC is linked to risk factors common in a western lifestyle, including obesity and carcinogenic substances found in alcohol and tobacco [10]. Men are more at risk to develop CRC than women, with a worldwide age-standardized rate (ASR) being 23.6 per 100,000 for males and 16.3 per 100,000 for females [11]. As the risk of development of cancers increases with age, and worldwide the elderly population is increasing, cancer incidence in general is expected to rise over the coming decades. Colorectal cancer cases are expected to rise faster than other cancers, as there is indication that incidence rates among younger people is increasing as well [12].

## 1.4 Clinical Management of Colorectal Cancers

### 1.4.1 Presenting symptoms and diagnosis

Most CRCs in Australia are diagnosed symptomatic (approximately 75%), although this number may fall due to implementation of routine screening through the National Bowel Cancer Screening Program (NBCSP) [13]. In this population screening programme, people aged 50-74 are tested every two years, using the iFOBT (immunochemical faecal occult blood test), which aims to identify microscopic blood in the stool [14]. Symptomatic CRCs may present at the general practitioner (GP) with symptoms of rectal bleeding abdominal pain, anaemia, weight loss and dyspepsia or ingestion [15].

Diagnosis is primarily conducted with a high-quality colonoscopy (the gold standard), which is performed following presenting symptoms as well as a positive iFOBT. Initial staging for colon cancer investigation uses imaging techniques such as a post-intravenous contrast-enhanced computed tomography (CT) scan of the chest, abdomen and pelvis, with additional Magnetic Resonance Imaging (MRI) or Positron Emission Tomography - Computed Tomography (PET-CT) to detect metastases [16]. Staging for rectal cancer, however, is primarily done with high-resolution MRI. A CT scan of the chest, abdomen and pelvis should not replace MRI, but can be performed as part of pre-operative staging, to assess for more distant nodal and metastatic disease.

### 1.4.2 Staging

CRC can be staged with various classification systems. Currently the Australian clinico-pathological staging system (ACPS) is Australia's preferred system, according to the accepted guidelines [17]. However, for international comparison, a TNM-based system is also recorded. A comparison of the classification systems ACPS, the related Concord substaging and the AJCC stage grouping based on the TNM-system in these stages are given in Table 1. The TNM-based system was introduced by the Union Internationale Contre Le Cancer (UICC) and the American Joint Committee for Cancer (AJCC) was introduced in 1986 and has been updated every few years since. The TNM-system uses coding for Tumour (size, or invasion depth of the tumour into the surrounding tissue), lymph Node infiltration and Metastases status. An overview of the descriptions of cancer stages in the TNM code-system is given in Appendix 1. The stage of CRC is critical to the decision-making in determining the treatment options. In this thesis, the ACPS version is used.

***Table 1: Comparison of Australian clinico-pathological staging classification with Concord and American Joint Committee for Cancer staging for colorectal cancer***

| ACPS | Concord substage | AJCC 8th edition (2017) | | | |
|---|---|---|---|---|---|
| | | Stage grouping | T | N | M |
| A0 | A1 | 0 | Tis | N0 | M0 |
| A | A2 | I | T1 | N0 | M0 |
| A | A2 | I | T1 | N0 | M0 |
| | A3 | I | T2 | N0 | M0 |
| B | B1 | IIA | T3 | N0 | M0 |
| | B2 | IIB | T4a | N0 | M0 |
| C | C1 | IIIA-IIIC | Any T | N1-N2 | M0 |
| | C2 | IIIA-IIIC | Any T | N1-N2 | M0 |
| D | D1 | 0-III | Any T | Any N | M0 |
| | D2 | IVA-IVC | Any T | Any N | M1a-M1c |

### 1.4.3 Prognosis

The prognosis for colon and rectal cancer patients are approximately the same, with an expected 5-year survival of 70% in Australia (2012-2016). This is a relatively good prognosis, compared to other cancers as lung (18.6%), liver (19.5%) and pancreas (10.7%), but worse than higher incidence cancers like prostate and breast cancer. In Australia, the mortality is below the worldwide average [18]. The 5-year survival has been rising from 50% in 1990 to 70% today, due to earlier detection and better treatment options. Even in higher age-groups, the 5-year relative survival is above 50%. In Australia as well as worldwide, males are affected slightly more than females (19 vs 14 per 100,000 ASR). Other demographic factors for higher mortality are indigenous status as well as living in lower socio-economic status (SES) areas [18]. Although patients with CRC overall have a good chance of survival, this is much lower in patients with metastasised CRC: the median survival is approximately 3 years, with 5-year survival of 20%. Early detection is key for a good prognosis.

### 1.4.4 Management of resectable tumours

**Colon cancer**

For early stage, non-metastasised cancer of the colon (Stage A-B), curative surgical resection of the primary tumour and anastomosis of the bowel is the preferred treatment [19]. Either an open approach or a laparoscopic approach can be considered, with no significant difference in survival outcomes, although a post-operative advantage in recovery time is found for the latter [19]. Approximately 70–80% of patients with newly diagnosed cases of colorectal cancer undergo curative resection. However, 40% of patients subsequently develop an incurable recurrent disease due to undetected micro metastases. To mitigate this risk, in locally infiltrated stages (stage C), adjuvant chemotherapy is standardly administered

for patients under 70 years, and sometimes for patients over 70 depending on their fitness. For stage II patients, the effectiveness adjuvant chemotherapy is not certain and is not part of standard care. Adjuvant therapy in the form of a biological agent (either bevacizumab or cetuximab) have not led to patients benefit and should not be considered [20].

**Rectal cancer**
Rectal cancer diagnosed at stage A can be locally or radically resected. For rectal cancer, an open approach is standard, but laparoscopic can be chosen in selected cases if the facilities and surgical expertise is available. Preoperative (neoadjuvant) radiation treatment (either short-course radiation treatment alone or long-course chemoradiation) is recommended for most patients with stage B and C rectal cancers, to reduce risk of local recurrence. For stages B/C, post-operative chemotherapy or radiation therapy can be considered as well [21].

**Recurrent CRC**
After resection of the primary tumour, cancers may recur locally, in 5–10% of patients, or systemically in 40% of CRC-patients [22]. Most patients with recurring cancer, present with pain symptoms and should be diagnosed with serum carcinoembryonic antigen (CEA), contrast CT scan of the chest, abdomen, and pelvis, and PET. Depending on the type of recurrence (local or metastatic), additional investigations might be necessary. A high-quality pelvic MRI is recommended for patients with locally recurrent rectal cancer. In locally recurrent colon cancers, pelvic exenteration is the primary treatment option, although no randomized controlled trial-study has been conducted. Locally recurrent rectal cancers can be treated with re-operative surgery as well, where neoadjuvant chemoradiation should be considered if radiotherapy was not used in the initial surgery. In the case of systemic recurrence, with resectable hepatic metastases, liver resection, possibly with adjuvant therapy, should be offered. In approximately 20% (10%–30%) of metastatic colorectal cancer (stage D, mCRC) patients, both the primary tumour as well as the metastasis can be treated with curative resection [23].

### 1.4.5 Management of non-resectable tumours

At the time of primary diagnosis of colorectal cancer, approximately 25% of patients present with synchronous metastases. Most patients present with symptoms in earlier stages of the disease or are found in screening practice. This is better for patient outcomes, as only a minority of mCRC patients are suitable for curative resection; approximately 20% [24]. The systemic care provided to mCRC patients consists of palliative treatment options, aiming to increase patients' survival and improve quality of life. Except in case the primary tumour obstructs or perforates the bowel, or when there is bleeding, there is no consensus in international guidelines whether surgery on the primary tumour site is beneficial, or that direct administration of chemotherapy is preferred [25]. Metastasis to the liver is most common in mCRC patients, with almost 50% of mCRC patients who will develop hepatic metastases over the course of their treatment. However, 80–90% of patients with liver metastases are not amenable for surgery and this results in liver metastasis being the dominant cause of death for patients with mCRC. Liver-directed therapies such as elective internal radiation treatment, radiofrequency ablation, hepatic arterial infusion of chemotherapy agents or trans arterial chemoembolization can be considered in context of a clinical trial, following up on US evidence [26].

Systemic first-line treatment for mCRC can contain a doublet or triplet form of chemotherapy as well as a biological agent (see Table 2). Single-agent chemotherapy is administered to patients that cannot tolerate a combination therapy. Triplet therapy FOLFOXIRI should only be administered to patients with good performance status and without significant comorbidities [27]. Increasingly, biomarker expression in CRC patients is evaluated to support decision making in the treatment regimen choice. Biological agents targeting pathways that are involved in tumour growth and spread are generally added to the chemotherapy regimen, and typically chosen depending on genetic mutations harboured by the tumour.

***Table 2: Overview of systemic therapeutic agents for colorectal cancer***

| **Form** | **Chemotherapeutic Agent(s)** | **Name** |
|---|---|---|
| Doublet | Fluorouracil (5FU) and leucovorin | FU/LV |
| Doublet | Leucovorin calcium (folinic acid), 5FU and oxaliplatin | FOLFOX |
| Doublet | Leucovorin calcium (folinic acid), 5FU and irinotecan hydrochloride | FOLFIRI |
| Triplet | Leucovorin calcium (folinic acid), 5FU, oxaliplatin and irinotecan hydrochloride | FOLFOXIRI |
| Doublet | Capecitabine plus oxaliplatin | XELOX/CAPOX |
| Doublet | Capecitabine plus irinotecan hydrochloride | XELIRI |
| **Target** | **Biological Agent(s)** | |
| VEGF | Bevacizumab | |
| EGFR | Cetuximab | |
| EGFR | Panitumumab | |

Biological agents targeting epidermal growth factor receptor (EGFR) or vascular endothelial growth factor (VEGF) in combination with chemotherapy are recommended in the first-line treatment for most patients. RAS mutations are routinely assessed to see if mCRC patients are eligible for therapies targeting EGFR. RAS wild-type tumours should be treated with anti-EGFR in combination with chemotherapy. As well, BRAF mutations should be assessed in mCRC patients, as it is considered a poor prognostic marker, with a low response to EGFR targeted agents. In clinical trial setting, emerging biomarkers such as HER2, HER3 and MET mutations can be assessed as well.

## 1.5 Motivation of the study

For improving health services, being able to analyse the efficiency and effectiveness of care is an important skillset. A large part of this skillset requires the researcher to adequately model the care provided to patients, including outcomes, service utilization and costs. Based on model-outcomes, recommendations can be provided to improve patientcare. We have seen a trend globally in adoption of value-based healthcare, a philosophy and payment methodology for healthcare management and policy, where patient-centredness is a central element [28]. In this approach, integrated care is organised (and modelled on) around a single disease type, resulting in integrated pathway models.

A systematic review into the value-based interventions on integrated pathways in oncology shows, outcomes are often lacking and are of variable quality when available. Despite promising early insights, the efficacy of these interventions in cancer remains unclear, partially due to lack of insights in the complete pathways [29]. To establish efficiency and effectiveness, more insights are needed in variation of care and in the costs and outcomes of the provided care. In the current modelling of the standards of care, optimal care pathways are developed with clinician consensus-based models, such as the optimal cancer care pathway for people with colorectal cancer by the Australian Cancer Council [30]. These models are foremost specialist consensus-driven, while one of the key objectives of value-based healthcare is to improve care by data-driven cost- and outcome research.

There are several limiting factors for good analysis of care and its costs for the integrated pathways. These limitations can be in collection, access, and analysis of the data related to these pathways. Hospital information systems routinely store data on the care provided, as well as on the clinical outcomes. The costs of different care activities provided to patients, is registered in some countries in their EHR or hospital information systems, but in some countries it is not due to the financial relations between patients, insurers, and medical professionals. Health Services Researchers often need to combine information from different sources to obtain insightful models.

Also, in diseases such as colorectal cancer, there is a complex system of care delivery, comprising of multiple care providers, from general practitioners to oncologic specialists, over a time span of multiple years or even decades. The registry of activities can be scattered across multiple IT-systems and combined with increasingly strengthened data-privacy laws, data-collection on all care provided, can be challenging.

Moreover, while clinical practice is based on the latest guidelines and recommendations, care provided to individual patients can differ. This difference can be based on patient characteristics, patient- or medical practitioner preference, availability of services and more. Not all practice variation can currently be explained, and it is desired that this unexplained variation is reduced. Modelling the care provided to patients, to understand clinical practice and make recommendations to improve, can become complex fast, if the care is very different in terms of type of care, costs of care and outcomes of care.

These combined factors make it time- and resource intensive to model which care is provided, analyse its costs and what outcomes are associated to it. By partially automating the modelling of the care process this time- and resource intensity can be reduced. A methodology providing a data-driven model of provided care over the entire pathway, with a quantitative substantiation of the quality of a model would benefit decision-makers in healthcare tremendously. This is where the novel data science subfield Process Mining can be of value.

Process Mining is a data-mining technique focussed deriving process models from low-level event data in electronic record systems. These process models provide a useful tool for evaluating performance of the system and analysing the variations in the underlying processes. In the context of health services delivery, interest has been growing into the research field of Process Mining, as the techniques used improve the level of detail in which care pathways can be analysed and may greatly reduce the amount of manual work needed for such analysis.

## 1.6 Study objective

The aim of this thesis is to investigate the care provided in the entire care pathway of CRC patients in an Australian setting, by applying process mining techniques on a multi-centre linked cohort of CRC patients. The interest is in the entire pathway of clinical practice, ranging from primary care diagnosis to first-line treatment in hospitals and includes diagnostics, prescriptions, procedures in the hospital. Additionally, the aim is to investigate the costs of care delivery in these pathways for groups of patients with CRC, using process mining techniques.

The study objective is formulated as:

> *Demonstrate the capabilities of Process Mining in the context of Health Services Research to map care pathways for patients with colorectal cancer, to quantitatively evaluate these pathways and to depict costs of care across the pathway.*

The research is intended as a Proof-of-Concept study, and while the interest in this study is specifically on care pathways of colorectal cancer patients, the methodology applied to this cohort should be generalisable to other cohorts and diseases as well. In the next chapter, an introduction to the field of Process Mining is given. A research question as well as more granular set of sub questions are provided in the end of this chapter, after introducing important Process Mining elements.

# 2 INTRODUCTION TO PROCESS MINING

This chapter describes the context of the data-science subfield of Process Mining. Process Mining (PM) is a data-mining technique focussed on structuring events, registered in so called event logs, into process models. The field of PM aims to generate relevant process structures from low-level event data, validate the quality of these process structures and use additional information to enhance or enrich these models. These process models provide a useful tool for evaluating performance of the system and analysing the variations in the underlying processes.

PM has been applied in various domains, from industry to IT and from governmental and financial audits to healthcare [31]. In the healthcare domain, process mining is mostly restricted to case studies, to assess its viability in a specific situation. Most of the studies are performed in more complex administration of care, such as oncology and surgery [32].

## 2.1 Process Modelling

Process modelling, specifically quantifying process structures and changes in these structures, starts with a formal definition of processes as a mathematical object. In its most basic form, a process is defined as a system that has multiple states, each available through execution of transitions. There exist multiple forms of mathematical notation of processes, such as Causal Nets (CN's), Event-driven Process Chains (EPC's), Yet Another Workflow Language (YAWL) models or Markov Chains [33]. The Petri net became the first model of a process to capture concur
rency [34] and is still the most common form in process mining, and most algorithms are based on this concept [35].

### 2.1.1 Petri Nets

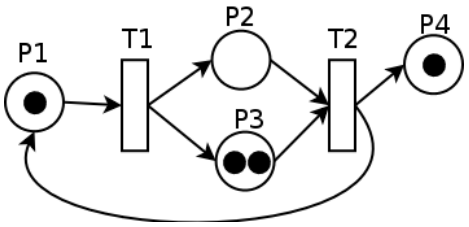
A Petri net consists of a multi-dimensional tuple, of at least 4 elements (P, T, F, M0). With these elements, a graphical notation can be produced, as seen in Figure 3. More elements can be added to extent the capacity of the model, for example in the context of Data-aware petri nets, where each transition or flow can contain another element as an additional dimension in the tuple. [36].

**Mathematical notation**

$P = \{P1, P2, \ldots Pn\}$
is a finite set of places
$T = \{ T1, T2, \ldots , Tn\}$
is a finite set of transitions
$F = (PxT) \cup (TxP)$
is a set of arcs (flow relations)
M0: P {1,2, ...}
is the initial marking

$P \cap T = \emptyset$ and $P \cup T \neq \emptyset$

**Graphical notation**




Circles: Places
Rectangle: Transitions
Lines: Flow relations
Black dots: Tokens in (initial) markings

***Figure 3: Mathematical notation and (sample) graphical notation of a Petri net.***

The *places* describe the before and after states, the *transitions* describe a change in the system and the *flows* represents arcs between states and transitions. The *(initial) marking* describes the number of tokens at the start configuration of the system. The petri net is a discrete-event model: it only changes when an action is performed, at which time, the token position is updated. When an action is performed, the transition consumes the number of tokens related to the number of arcs going into the transition. It then returns one token in all places connected with arcs, going out of the transition. The structural similarity of Petri nets can be quantified by comparing differences between the available places, transitions, and flows.

## 2.2 Process mining

Process mining aims to form a bridge between classical data mining and business process analysis. The first papers on the abstraction of a process from events stored in data management systems are from 1998 and 2000 [37], [38]. These works became the basis of the alpha-miner, the first discovery algorithm. After this, more perspectives to PM were introduced and more subfields appeared. This resulted in perspectives looking at the order of process activities and interaction of people, and the three main subfields of process mining: process discovery, conformance checking, and process enhancement.

### 2.2.1 Perspectives

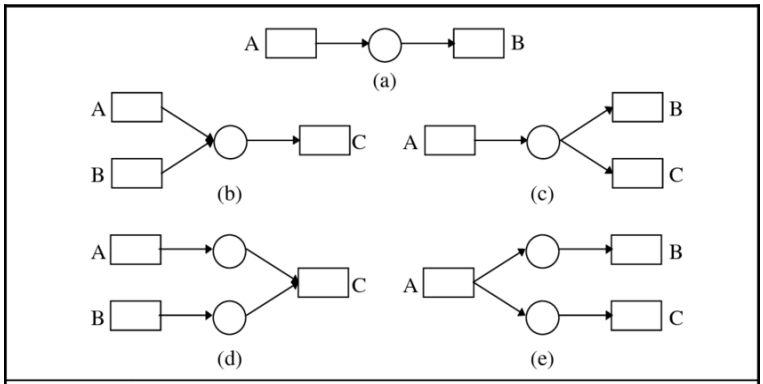
A process can be evaluated from multiple perspectives. In PM, the focus on the sequencing of events, called the *control-flow* perspective, was the first and still is the most researched perspective [39]. The concept of using algorithms to derive relations between events that happen in sequence had originally been named process mining and has later expanded to include the other perspectives [38]. PM can focus on the different uses of resources, as done in the *organisational* perspective, or on the people collaborating on the sequence of events, called the *social* perspective. When focussed on frequency and duration of activities, this is called the *time* perspective.

When taking a control-flow oriented approach to a process, one evaluates the sequence of steps in a process that comprises an action to reach a certain goal. It therefore has a clear start- and endpoint, with actions connecting them. Also, during the execution of the process, the system cannot be in multiple states at the same time. Using PM to derive Petri nets from a business process will lead to a workflow model [40]. Ideally, a workflow model is denoted as a special subclass of Petri nets called workflow nets (WF-nets) which have the *soundness* property. Soundness guarantees that a process can formally be finished, and this property is essential for conformance checking and process replay. However, not all discovery algorithms guarantee sound models, so a workflow model does not necessarily result in a WF-net.

### 2.2.2 Process Discovery

Discovery entails the automated derivation of relations between events in a dataset, which are stored in an event log. Such an event log consists of records of data, which capture one or more activities (an *event*) in a single instance of a process (a *case*) and was registered at a certain point in time (containing a *timestamp*). The sequence of events from a single case is then called a *trace* [41]. Many organisations routinely collect event data in the form of Enterprise Resource Planning (ERP), financial transactions, or electronic health records (EHR). A mining algorithm evaluates all the traces in the log and derives relations between consecutive events in the trace.

Discovery algorithms are able to derive at least 5 relations, given in Figure 4. Direct successive relations (Figure 4a), causal (conflicting) relations (Figure 4b/c), parallel or concurrent relations (Figure 4d), and choice relations (Figure 4e). With these relations, basic Petri nets can be generated, creating before and after states P for each transition T and using the relational aspects found into the flow relations F between the places. More advanced algorithms can derive more complex relations, such as (self-)loops with repeating activities or longer loops.




| | |
|---|---|
| Sequence: | (a) activity A must be completed before activity B can start |
| Conflict: | (b) C can start after either A or B have completed<br>(c) either B or C can start after completion of A |
| Concurrency: | (d) A and B can be executed independently (possibly in parallel), C cannot start until both of them have occurred<br>(e) once A has occurred, B and C can be executed independently (maybe parallel) |

***Figure 4: Basic relations between activities in a Petri net-based process model.***

Since the introduction of the alpha-miner, extensions have been introduced [42]–[44], as well as several new algorithms, including the heuristics miner [45], genetic miner [46], and the fuzzy miner [47]. All these algorithms have advantages over the basic alpha algorithm, but all also have the disadvantage that they do not guarantee soundness in their models. In 2013, the inductive miner (IM) was published by Leemans et al [48]. IM provided improvement in some aspects over the previous miners as it can cope with infrequent behavior and large event logs in polynomial time, while guaranteeing soundness [49].

### 2.2.3 Conformance checking

The process mining technique to assess the quality of a process model is called conformance checking. The technique takes as an input a process model and an event log and returns a set of differences between the behavior captured in the process model and the behavior captured in the event log. There are four quality dimensions for comparing model and log: (replay) *fitness*, *simplicity*, *precision*, and *generalization* [50].

*Fitness* quantifies how much of the observed behavior of the event log is captured by the model; *precision* quantifies how much behavior exists in the model that was not observed in the event log; *generalization* quantifies how well the model explains unobserved system behavior; and *simplicity* quantifies the complexity of the model. These quality metrics are often a trade-off: increasing one may lead to deterioration of the others. While there are other methods for evaluating the conformance of process models, such as rule-based conformance checking [51] and token-based conformance checking

[52], alignment-based conformance checking has been the standard since its introduction in 2014 [53].

An alignment is a computation which gives the equivalence of a trace and a model. It requires (a set of) traces and a model as input, and it calculates alignment , $\gamma i$, for each trace , I, in the set. $\gamma i$ consists of sequence of pairs that refer to an event from a trace and a transition in the model, or $\gg$ elements indicating deviations. A visual example is given in Figure 5. When the upper and lower part of a pair is the same, there is a synchronous move. When a transition in the model fires, but no activity is found in the log this is called move-on-model ($\frac{\gg}{A}$). When an activity is found, but the model cannot fire a matching transition, this situation is called a move-on-log ($\frac{A}{\gg}$). In the optimal alignment algorithm, for each trace, the problem is projected as the shortest path-problem and solved using the $\mathbb{A}^*$-algorithm [33], resulting in optimal alignment $\gamma^*$.

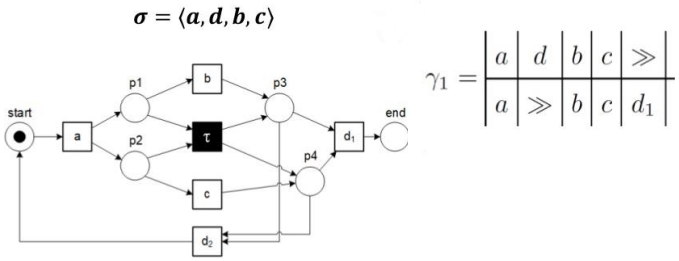



***Figure 5: Concept of aligning event trace and process model (left) and a computed alignment $\gamma^1$, where the upper row is the trace and the lower the model execution***

When adding the alignments as a decoration to a Petri net, for each transition, or event in the model, the number of synchronous moves as well as moves on log are displayed. For an activity, the alignment for the activity is then computed as the fraction of synchronous moves divided by the total number of synchronous moves and moves on model. With this fraction, we can find at which point in the model, the most deviation is compared to the log.

For each alignment $\gamma^*$, the *fitness* of the alignment is expressed as one minus the minimum number of moves needed divided by the total number of pairs in $\gamma^*$. The overall fitness can then be expressed in multiple ways: as the percentage of perfectly aligned traces ("Full fitness"), or as the average of the fitness per trace ("Trace-fitness"), or the percentage of all minimum moves to all pairs ("Log-fitness").

*Precision* is calculated with deviating behavior from the model to the behavior observed in the log: for each trace in the log, a maximum *anti*-alignment is computed, which is a firing sequence of the model resulting in the maximum number of moves to the trace. The number of fired transitions is bound by the length of the original trace.

*Generalization* is more challenging to calculate and relies on an estimation of the probability that a new observation of an activity in a trace can be properly explained by the model. The metric takes the law of large numbers into account: if most activities in the log are observed a large number of times, the probability that the next activity in a trace is an unseen one, decreases. The opposite is true as well: the probability that a new activity that cannot be explained by the model is found will be high if most activities in a log are unique.

*Simplicity* penalizes models that are more complex without additional value. Unlike the other conformance metrics, simplicity can be measured without considering observed behavior of process executions. When considering several Petri nets with/allowing the same set of traces, the simpler Petri net is the one with fewer number of duplicate/invisible transitions and implicit places. The simplicity of a Petri net is measured based on the number of activities that it represents and the number of control-flows it has.

Combined metrics, such as a weighted average for the four dimensions, and the f-score, which weighs the fitness and precision, are used as well. An overview of the metrics and what aspect of model quality they address is given in Table 3.

***Table 3: Overview of the conformance metrics and which aspect of model quality they evaluate.***

| **Metric** | **Model quality aspect** |
|---|---|
| Fitness | Measures how many traces or activity sequences in the total of all traces in the reference log could be correctly displayed by the model. |
| Precision | Measures how many additional traces or activity sequences the model could display, from activities that were observed in the reference log |
| Generalization | Measures how many additional traces or activity sequences the model would be able to display, from activities that were *not* observed in the reference log |
| Simplicity | Measures how many model components were needed to display the number of unique activities in the reference log. |
| F-score | Balanced average of the Fitness and Precision |
| Average conformance | Balanced average of the four main conformance metrics |

### 2.2.4 Process Enhancement

Process Enhancement is most commonly defined as the extension or improvement of an existing process model using information about the actual process recorded in some event log [54]. Two forms exist. Firstly, *repair*, where commonalities and discrepancies found in the control-flow perspective (found with conformance checking) are used to improve the model. Secondly, *extension*, which adds information from other perspectives, such as the organisation ('resources-perspective'), the frequency and timing of events ('time-perspective') or the properties of the case ('case-perspective') to the process model, to improve the insights of the model.

In the context of cost-analysis, an extension can be introduced from the cost-perspective [55]. This perspective, in contrary to the resources-perspective focusses on the cost drivers for the execution of a transition rather than on the attributes of an event. Some efforts are directed towards cost-perspective based process enhancement [56], [57]. However, the cost-perspective is not a well-researched perspective and literature reviews on PM do not identify the perspective at all. [32], [58]. Analysing costs in the process mining cost-perspective, as an extension of the process model for a care pathway, can provide a novel way to visualise and compare care provided to patients and is generalizable to a broader context, including outside the scope of healthcare.

## 2.3 Terms and definitions for Process Mining

In the field of PM, multiple definitions and notation forms exists for the various algorithms, process models, and model components. The Institute of Electrical and Electronics Engineers (IEEE) Task Force on Process Mining has the goal to promote the research, development, education, implementation, evolution, and understanding of process mining, and released the 'Process Mining Manifesto' where the basic propositions of the field are described [59]. These authors are responsible for some of the earliest research in this field and the notation they introduced are most common. Throughout this thesis, a set of terms and definitions that is based on Van der Aalst's '*Process Mining: Data Science in Action'* will be used in the context of PM, with some extensions to hierarchical models described by Yang et al. [41], [60].

When we define a process as (a series of) actions or activities to achieve a certain goal, we can store process data in an event log $\boldsymbol{L} = \{1,\ldots, n\}$ where each element n represents one process case $\boldsymbol{n}$ (see Table 3).

***Table 4: Minimal required information for an event log***

| Case-ID | Event-ID | Timestamp |
|---|---|---|
| **00001** | Activity A | 2008-02-01 20:00 |
| **00001** | Activity B | 2008-02-01 21:30 |
| **00001** | Activity C | 2008-02-02 8:30 |
| **00002** | Activity A | 2008-02-01 20:30 |
| **….** | …… | ….. |

***n***

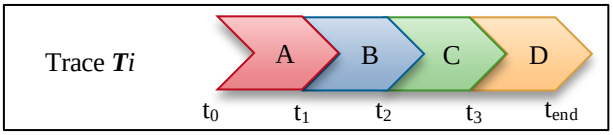



***Figure 6: Trace example***

One process case consists of two elements $\{i,\boldsymbol{T}i\}$, as it is indexed with a unique case-identifier $i$ and consists of the activity trace $\boldsymbol{T}i$, a tuple which has an event-identifier and a timestamp. A unique activity trace (see Figure 6) can be represented as $\boldsymbol{T}i = [a1(i),\ldots,ak(i)]T$, where $aj(i)$ is the $j$-th activity with name $a$ (out of $A$ possible activities), in trace $\boldsymbol{T}i$, sorted by activity start time, and $k$ is the trace length (i.e., number of performed activities for this case).

In the most basic form, when applying a process discovery algorithm on log $\boldsymbol{L}$, a process model $\boldsymbol{\lambda}$ can be obtained. A model can be obtained from the complete log $\boldsymbol{L}$, but also from a subgroup of traces in the log $\Phi$ consisting of individual traces $\{\boldsymbol{t}_1, \boldsymbol{t}_2,\ldots,\boldsymbol{t}_i\}$. A subset of certain traces may lead to more homogenous groups and thus more homogenous or less complex process models.
Additionally, models can be obtained from parts of entire activity traces or a *phase* in the process. In this case, each trace $\boldsymbol{T}i$ can be split in $\boldsymbol{T}i^* = [aj(i),\ldots,an(i)]T$ , where trace $\boldsymbol{T}_i^*$ is a subset of $\boldsymbol{T}_i$, only taking activities $j$ through $n$ into account. Then from log $\boldsymbol{L}_{\mathrm{ID}}$ or subset $\Phi_{\mathrm{ID}}$, containing all split subtraces $\boldsymbol{T}_i^* = [aj(i),\ldots,an(i)]T$, process model $\boldsymbol{\lambda}_{\mathrm{ID}}$ can be derived, where ID is the index of the phase in the set of phases defined by the subsetting method of the partial trace.

For useful and adequate process mining, there needs to be a definition of what constitutes an activity $ak(i)$, on an adequate level of granularity $h$. A process model might consist of multiple levels of granularity, in which individual events can be clustered to a group. For example: the start, pause, continuation and the ending of an event may be registered separately, but can be grouped to one instance of that event. Also, within the context of the field PM is applied to, clustering may be done on the basis of grouping related to that field, such that activities that are part of a larger subset of activities may be clustered together. A visual example of this hierarchical process structure is given in Figure 7.

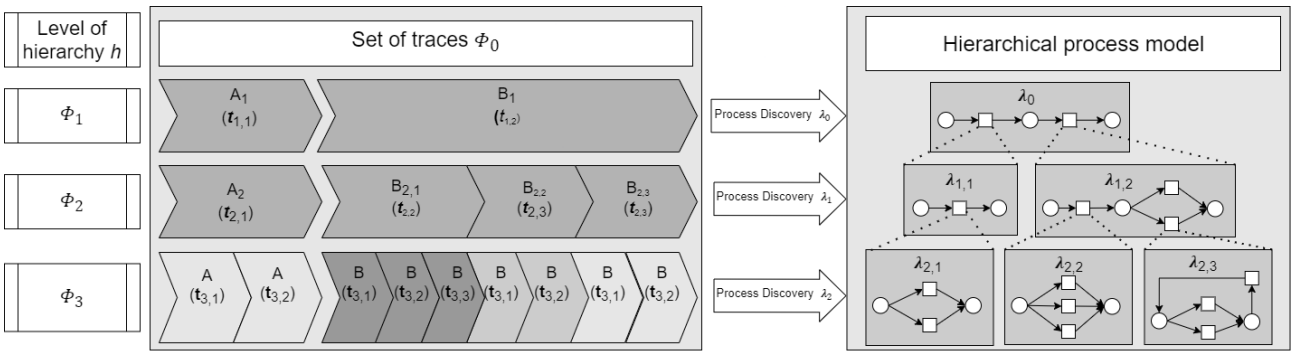



***Figure 7: Discovery of process models in a hierarchical setting***

When applying PM, you can evaluate the trace on the clustered level, or the unclustered level. In this case process model $\boldsymbol{\lambda}_{\mathrm{h}}$ can be obtained from $\boldsymbol{L}_{\mathrm{h}}$ or $\Phi_{\mathrm{h}}$, where h is the granularity level on which $\boldsymbol{Ti}$ is evaluated.
Numeric attributes such as costs or waiting times can then be aggregated or decomposed to this desired level of granularity. Previous research has shown both supervised methods of designing hierarchical structures in processes, as well as automated algorithm-based methods [61]. Because of the specialized nature of care pathways, as well as the observed complexity of care delivery, with frequent repetitions of specific steps, supervised design of outlines of the carepathways can be beneficial, because it restrains the models allowed behavior and with it, the computational requirements to run algorithms on the process models.

## 2.4 Software used for Process Mining

Process Mining software comes in various academic and commercial packages, with each their specific advantages. The most used software package is ProM, the academic process mining suite. It is a Java-based platform, developed by the Process Mining Group from Eindhoven University of Technology [62]. ProM contains more than 800 extensions, all publicly available. While the performance of this language is high and most computations can be performed in real-time or in relatively short timeframe, the development of the platform is limited to computer science researchers. The platform has a steep learning curve and requires the user to be able to code in Java. The focus for users is more on the application of the different extensions and use ProM as a 'Workbench' as the Process Mining Group calls it, rather than the development of new tools.

Process Mining software is available in a fast-growing number of commercial tools, which aim to give data analysts the tools to apply process mining on their own datasets and use the results to drive their decision-making. Fluxicon's Disco, Celonis Process Mining suite and Apromore are among the larger software suites for PM in organisations [63]. These suites are known for their user-interface and many low- or no code extensions, but for financial, security and stability reasons, the developers do not allow modifications to their software. Scientific experimentation with custom code with these suites is therefore not an option.

Both for R, and Python, two of the worlds most used languages for Data Science, free frameworks for process mining have been developed. For R, it is the bupaR (Business Process Analysis for R) suite and for Python, PM4PY (Process Mining for Python), where it should be noted that some components in the bupaR suite are built on the underlying framework of Pm4py. Both suites have been developed for faster algorithmic customization and scientific experimentation [64], [65].

The choice which software to use when starting a process mining project, is strongly dependent on the data quality and the level of programming proficiency required for the projects research problem. If there is data available of relatively good quality, from which conclusions on the process itself are desired, the standard packages can be a good choice. If the objective is to try to develop new PM algorithms that Computer Scientists can compare to the existing ones, the open-source program ProM and language Java is likely the best option. For Proof-of-Concept studies, with data that requires additional wrangling, either the R-based or Python-based frameworks are preferred.

## 2.5 Literature research

Process mining is a relatively young discipline, having originated from research in the late 90s and early 2000 at the University of Eindhoven [37], [38]. In the first years, the focus was mainly on the development of suitable discovery algorithms, with the development of the heuristic and genetic mining algorithms [45], [46]. The initial tooling, ProM, was launched in 2004 and continued to grow to be the most used software platform for process mining, where researchers could develop their algorithms and share them as a plugin [66].

From 2007 onwards, the discipline matured and started expanding outwards, applying the developed algorithms on various domains, including finance, government and industry [31]. Process Mining set foot in the healthcare domain in the early 2010s, driven by the First International Business Process Intelligence Challenge (BPIC'11), where the dataset used in the first healthcare related paper by Mans et al. [67], was provided with the challenge to apply interesting use-cases and techniques on a medical dataset.

In 2016, Rojas and all provided a literature review of clinical case studies using Process Mining in the healthcare domain over the period until then, identifying several common aspects for comparison, including methodologies, algorithms or techniques, medical fields, and healthcare specialty [32]. This extensive study is the go-to paper to understand the potential of Process mining as a tool to analyse complex care provided to patients. According to Rojas, oncology was one of the most researched topics within Process Mining in healthcare and this is still the case in 2020 [68].

Process Mining proved to be a useful tool in the healthcare domain for evaluating the compliance with guidelines of actors in the healthcare system [69], for analysis of resource-usage and collaboration of physicians [70] and for analysis of common workflows or bottlenecks in the system [51], [71]. The most found research was on conformance checking to identify deviations from clinical protocols, such as by Yang et al. which proposed a methodology for the automatic detection of deviations from the established protocol in a hierarchical workflow model, based on the clinical airflow resuscitation process [60]. Recently, another case study by Sato et. all applied this on different levels of granularity of the process, in a case study on bariatric surgery [72].

Up until 2015, according to Rojas[32], Process Mining research in the healthcare domain focussed mainly on individual treatments, individual medical organisations, or individual registries of electronic health records. Process Mining research with use of linked data from multiple sources has yet to be completed. Using linked data requires thorough anonymization of patients and sector wide naming conventions, suited for process mining, which are identified as one of the main data challenges in the medical domain by Mans et al [73].

While there is research in automatically creating models for clinical pathways, with other techniques than Process Mining, the research in this area is limited. In a notable study, Cho et al. proposed a non-PM algorithm for developing data-driven clinical pathways using electronic health records, applying it on to case studies, one for total laparoscopic hysterectomy and rotator cuff tears [74].

In the context of Value based healthcare, Ibanez-Sanchez et. al aimed to perform an analysis of the ways in which Process Mining techniques can support health professionals in the application of Value-based technologies. Their research demonstrated how Process Mining technology can highlight the differences between the flow of stroke patients compared with that of other patients in an emergency [75]. In the context of health services research, Yampaka et all combined process mining with queueing theory, developing a model that was suitable for analysing control-flow and time performance in health service domain [76].

As recently as 2020, Helm et al. proposed a set of standardized terms in clinical case studies for process mining in healthcare [77]. In the same year, Martin et al. proposed a set of recommendations for enhancing the usability and understandability of process mining in healthcare, including setting up a benchmarking study to identify the most suitable process modelling language to visualize the output of control-flow discovery algorithms in healthcare and developing techniques to handle data quality issues in healthcare event logs [78].

Ghasemi et al. [68] concluded from their systematic literature review that Process Mining applied in the healthcare domain is expanding rapidly. We see multiple journals that issue calls for papers for Special Issues, including the Journal of Biomedical informatics and the International Journal for Information Retrieval Research [79], [80].

It can be concluded that the application of process mining in a healthcare environment will continue to grow and become a full-fledged domain of research. The research will likely focus on complex care systems, encompass discovery, conformance and enhancement algorithms, and will likely be combined more with other fields such Data Mining and operations research. Challenges lie within standardization of nomenclature of the data, as well as developing more automated software, to reduce the amount of time data preparation takes.

## 2.6 Research questions

Returning to the study objective from section 1.6:

> *Demonstrate the capabilities of Process Mining in the context of Health Services Research to map care pathways for patients with colorectal cancer, to quantitatively evaluate these pathways and to depict costs of care across the pathway*

There are three distinct elements within the objective that are of interest to explore.
First, we would like to study how care pathways of colorectal cancer patients can be constructed using (one of) the discovery algorithms and to evaluate the quality of the resulting models. The quality metrics found with conformance algorithm can then be used to evaluate equivalence, by cross-examining the cohorts.
Secondly, resulting from the first question, using the conformance algorithms, we should be able to find which patient groups are more or less conformant to the pathway of the entire population. With this information, we can find the characteristics of the patients that are most conforming to the main pathways.
Lastly, we would like to go beyond the control-flow perspective and analyse the derived pathways from the cost-perspective, in order to find disparities in costs between the subpopulations. To do this, the models need to be enhanced with information on the costs, displayed in the visualizations and compared between the subpopulations. The total costs can then be calculated by aggregating from individual events in each of the trace.

Combined, the main research question is then formulated as:

*"How can Process Mining be applied to derive care pathways and analyse the costs of care provided to CRC patients in these care pathways?"*

With the following subquestions:

> *How can data-driven models for care pathways for CRC patients be derived with process mining techniques and how can these care pathways be evaluated?*

> *What are the characteristics of the patients going through the (main) CRC pathways?*

> *How can the total costs of the (main) CRC care pathways be calculated and evaluated for specific subgroups?*

In the next chapter, the methodology to answer these questions is described, followed by a chapter where this methodology is applied on a multicentred linked dataset of CRC patients.

# 3 METHODOLOGY

This chapter introduces the set of methods used to investigate the research problem described in the previous chapter. It provides the software, algorithms and parameters used in the thesis and specifies a method to evaluate costs accumulated in complex process models and describes a workflow to select a cohort, pre-process the data to event logs, discover and check the process models for carepathways and enhance these with costing.

## 3.1 Software, algorithm, and mining parameters

For this research project, due to the Proof-of-Concept nature of the study and the desire to add additional code to the algorithms, a combination of tools from R and Python are used. Pm4Py in native Python (version 3.8) is used for discovery of process models with the inductive miner and applying conformance algorithms. Part of the analysis is performed making use of R (R 3.6.3) using the bupaR process mining suite [81] (version 0.4.0.9000) and the Process Mining for Python (PM4PY) extension, running on Python (version 3.8). The enhancement part of the analysis is performed on process maps, containing visualization elements from packages ProcessmapR, and ProcessAnimateR. Visualizations are created using both these packages, as well as GraphViz.

The algorithms selected for PM are partially based on what is available in the chosen software. While the academic ProM software package contains a larger number of (experimental) algorithms, the commercial packages and the frameworks in R & Python have limited options. There are three process discovery algorithms available within the PM4Py and bupaR process mining suite; Alpha, Heuristics Miner and Inductive Miner, as well as the Direct-Follows Graph (DFG), which is not classified as a discovery algorithm. The latter is an algorithm to construct a graph, where each of the nodes are available and for each time two activities are observed after each other, an arc is added. This form of discovery is not suitable for conformance checking. In the bupaR suite, not all versions of the conformance algorithms are available, so conformance checking is performed solely in Python. The entire analysis was performed without parallelisation and was executed on a laptop (Lenovo 470T, Intel Core i5-7200U, 2.5 Ghz).

**Discovery algorithm**

To answer the research questions, one of the three available discovery algorithms was chosen for all experimentation. The alpha miner algorithm is one of the most basic algorithms for Process discovery and is unable to discover loops of length one, of length two or more and non-local dependencies. [82]. Especially the former two are problematic in discovering processes that might contain activities that are being executed multiple times. A similar problem occurs with the Heuristic miner algorithm, that cannot handle to many different events and cannot identify longer loops. Both Alpha and Heuristics miner algorithms result in spaghetti-line models as shown in Figure 8. While mathematically sound, these models cannot be interpreted by humans easily. The Inductive Miner algorithm was the only algorithm which could identify loops and produced interpretable models, see Figure 8b.

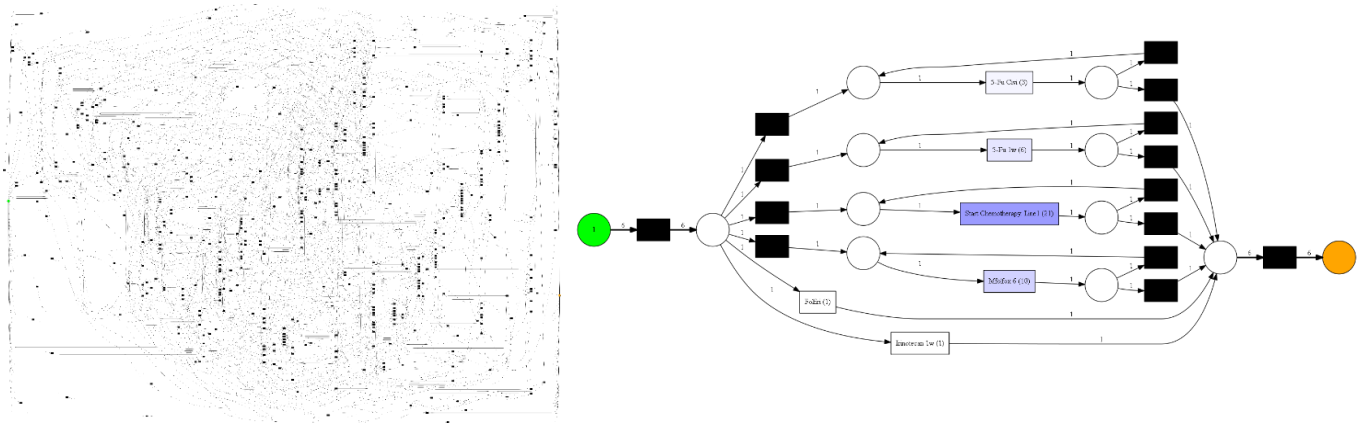


***Figure 8: Spaghetti-like model, derived with the Heuristics Miner algorithm (left) compared to visually interpretable model, derived with Inductive miner (right)***

Three versions of the Inductive miner algorithm are available, the basic IM version, the IMf (infrequent) behavior version and the IMd (directly-follows) version, where the second produces a more precise model at the expense of some very infrequent behavior, and the last considers the DFG first, to increase performance, but loses the replay fitness guarantee. As replay fitness is essential for alignment-based conformance checking, the latter was not an option. The
IMf version was eventually chosen, as the computation time was 10% faster, without observing a difference with the IM version on the test data.

**Conformance algorithms**

For conformance checking, more algorithms are used at the same time, as model quality is a trade-of between the different quality dimensions: Fitness, Precision, Generalization and Simplicity. Within PM4PY, the algorithms for fitness, precision and generalization have two versions implemented: The token-based version and the alignment-based version. The alignment-based version is not an approximation and takes approximately 30% more computational time, and it was observed that the outcomes of the fitness were a few points higher using the test data. The alignment-based version is chosen, as the model quality was regarded worth the computational time increase. The simplicity calculation is not dependent on an event log and the algorithm only has one version. Next to these general algorithms, the average conformance (mean of the four values) and the f-score (balance between fitness and precision) are calculated as well.

## 3.2 Cost aggregation in Process Mining

For costing the elements of care in the pathway, there needs to be an aggregation step of costs of individual components. The PM4PY package does not yet contain a possibility to aggregate custom variables over components in the Petri nets. As the objective is to enhance the process maps with additional information regarding the costs of each of the executed process steps, an extension is written that is able to aggregate the value of a custom defined numeric value, over all aligned traces to a Petri net. This function takes in four objects: an event log $\boldsymbol{L}_h$ which contains a set of traces to be aligned to the Petri net, the discovered Petri net itself, as well as the start- and final marking $M_0$ , $M_f$ and an aggregation function, such as the median, the mean or the sum. A visual representation is given in Figure 9 and the pseudocode is provided in Table 5. The initialization step of the algorithm initializes an empty list, **O,** of numeric values, where the current cost value will be stored. Then first, from the event log, $\boldsymbol{L}_h$, the set of traces, $\Phi_h$, is stored and a list of all the activities or transition, $\mathbf{T_{H,I}}$, in the model, $\lambda_H$. Secondly, For each level of hierarchy evaluated, each trace σ in $\Phi_h$ is aligned to model $\lambda_H$. Then, for each activity $aj(i)$ in the trace σ aligned to transition, $\mathbf{T_{H,i}}$ , the associated cost value is aggregated by the specified aggregation function and stored in the initialized list **O.** When all traces have been completed, the list, **O**, is concatenated to the list of transitions $\mathbf{T_{H,I}}$ in the model $\lambda_H$, resulting in an annotated model.

The main benefit of this way of modelling costs, is that it takes into account whether an activity was actually present at that location in the original model. Costs of activities that were not present in the original model, are an 'unexplained' value. When this value is large, compared to the entirety of the costs of all the patients, you have a model that might have a good fit with your data, but it lacks the elements that drive up the costs the most.

***Table 5: Pseudocode node aggregation in Petri net***

| | |
|---|---|
| **Input:** $\boldsymbol{L}_h$, $\lambda_H$, $M_0$ , $M_f$ , $\boldsymbol{f}_{aggregation}$<br>**Output:** $\lambda_{H,annotated}$ | |
| Step 0 Initialise | 1 Initialise list ***O*** with k components where k = number of unique transitions $T_H$ in the petrinet $\lambda_H$ |
| Step 1. Subset Traces from log | **For** each k in ***O:***<br>1. Let $\Phi_H = \{\boldsymbol{t}_H1,\boldsymbol{t}_H2,...,\boldsymbol{t}_Hi\}$ denote the set of traces in $\boldsymbol{L}_h$ of all case *i* in I on aggregation level H<br>2. Subset from $\boldsymbol{L}_h$, all $\mathbf{T_{H,i}}$ in $\Phi_H$ associated with $\lambda_H$. |
| Step 2. Align traces on each Level | **For** each h in ***H***:<br>1. Let $A_H$ denote the set of aligned traces of all case *i* in I on aggregation level H<br>2. $A_H$ = Alignments $(\Phi(V)_H, \lambda_H)$ )<br>3. Let $M_H$ denote the set of activities in the aligned traces $A_H$ associated with $\lambda_H$ of all case *i* in I on aggregation level H |
| Step 3. Aggregate node | **For** each unique $T_H$ in $M_H$:<br>1. Let $\boldsymbol{O}_k$ denote the total value of custom attribute on transition $T_H$<br>2. $\boldsymbol{O}_k = \boldsymbol{f}_{aggregation}(T_H)$<br>next $T_H$ |
| Step 4. End | Next h<br>Next k |
| Step 5. Add decoration to net | **For** each $T_H$ in $\lambda_H$:<br>**For** each $M_H$ in $A_H$<br>**IF** $T_H = M_H$<br>$\lambda_{H,annotated} = = \boldsymbol{f}_{add\_to\_net}(\boldsymbol{O}_k, \lambda_H)$<br>End IF<br>next $M_H$<br>next $T_H$ |
| Step 6. Output: | **return** $\lambda_{H,annotated}$ |

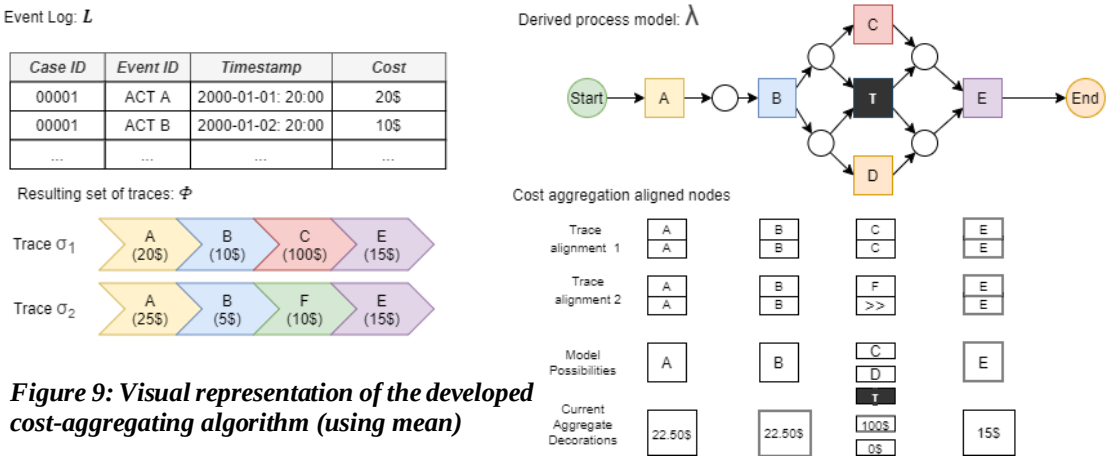



***Figure 9: Visual representation of the developed cost-aggregating algorithm (using mean)***

## 3.3 Process Mining workflow

To standardize the components of PM techniques that will be applied to patient data, a workflow is described. These steps are clustered and related to the three questions in Section 2.6. These questions are specified to the context of CRC, but the steps in the workflow should be generalizable to other types of diseases as well. The bold part of the text should then be replaced with the patient population of interest.

*1. How can data-driven models for care pathways **for CRC patients** be derived with process mining techniques and how can these care pathways be evaluated?*

**1.1.** Identify and select cohort of interest

**1.2.** The first step is to identify the cohort of patients that is of interest in the model, by finding all unique patients in the different datasets. Then these unique patients are linked across different datasets, by joining their unique identifiers. A table of each unique patient, with all patient-related attributes, should be a result from this step in the workflow. Each patient can then be linked with the information in all other datasets.

**1.3.** Classify phases of care delivery

**1.4.** The second step is related to the (possible) hierarchical relation between steps in the care process. It is needed to choose on which level of aggregation the modelling and analysis takes place and to do so, the phases of care are described. This process is based on insight in the domain of analysis, as the choices are made subjectively. The attributes of a care activity, such as the name and the date it took place, are then clustered into groups based on this phase. For example, all elements for CT-scans and related activities in the dataset are grouped into a 'Diagnosis' phase. Multiple clustering steps can be made, resulting in higher order phases.

**1.5.** Transcribe elements registered in a phase to activities and convert this to an event log.

**1.6.** The third step is to convert the data to the suitable format for process mining, the event log. In the table-structure of the data, the elements in a phase can consist of one or more descriptions with one or multiple dates associated to them. The same as the previous step, this process is based on insight in the domain of analysis. For example, in hospital care, when treating the patient with chemotherapy, the records can contain dates for each cycle of chemo-distribution to a patient, with a list of drugs related to the regimen and their dosage. This can then be transcribed as one activity of regimen X, with a start-date on the first distribution of that regimen and one activity with the same name on the end-date. The costs are then added as an additional attribute, summing all costs of the drugs at their described dosage. This example can be transcribed as a cycle of that regimen, where a list of activities is produced with the dates of each distribution. The conversion to an event log object is an automated step in all PM-software packages.

**1.7.** Apply a discovery algorithm on the entire event log

**1.8.** The fourth step is to apply the selected discovery algorithm, on the entire event log and that of each single phase. Using the chosen software and parameters, this step is automated and results in a Petri net, modelling the entire pathway or phase respectively. The model can then be visualized for human inspection.

**1.9.** Apply conformance algorithms on the resulting petri nets

**1.10.** The fifth step is to assess the quality of these base-line models, by applying the selected conformance algorithms on the Petri net and the same log it was derived from. This results in the base-line quality-metrics for the model of the entire phase.

**1.11.** Filter the event logs of the different phases based on selected characteristics

**1.12.** Based on selected patient characteristics of interest, such as the patient's gender, age, location, or any other attribute that is captured in the data, the log can be filtered resulting in a subset of patients ($\Phi_H$**)**. This step is needed to create comparator groups that are of interest.

**1.13.** Apply discovery and conformance algorithms to patient subsets based on selected characteristics

**1.14.** The seventh and last step is to sequentially apply discovery and conformance algorithms to the filtered subsets, resulting in a list of process models and their corresponding conformance metrics.

*2. What are the characteristics of the patients going through the (main) **CRC** pathways?*

**2.1.** Align the subpopulations event logs to the petri net of the whole pathway and pathway per phase
**2.2.** The obtained subsets can also individually be compared to the derived process model from step 1.4, by aligning this event log to that resulting Petri net. From the obtained alignments, the conformance measure of the subpopulation to the model can be obtained, as well as frequencies of patients having each individual activity in the model.

**2.3.** Compare conformance metrics between different subpopulations based on selected characteristics
**2.4.** Compare frequency fluctuations between different subpopulations based on selected characteristics

**2.5.** Display characteristics of the patient groups with a high conformance to the pathway and a large frequency of occurrence
**2.6.** From the resulting comparisons, the characteristics of the patients that align or go through the main pathways can be displayed.

*3. How can the total costs of the **(main) CRC** care pathways be calculated and evaluated, using process enhancement from a cost-perspective?*

**3.1.** Cost individual care activities in each phase
**3.2.** The care activities that are taken into account, must have a cost associated to them, which is a process that requires additional information to the datasets, either through insurer reimbursement data or specific cost associated schemes for reimbursement. This step requires domain knowledge as well as enough information in the data to relate price data to.

**3.3.** Create density distribution of all costs accumulated over the pathway and per phase
**3.4.** To compare the distribution of all costs over the entire pathway, a density distribution or histogram can be created from the costs associated to all activities in a selected phase. This can be done for each subpopulation of interest.

**3.5.** Apply cost aggregation function on Petri nets of subpopulations
**3.6.** The cost aggregation function, as described in paragraph 3.4 can be used to aggregate costs of a certain subpopulation over each node in the Petri net. This results in a decorated Petri-net with aggregated costs added to each transition or each activity in the Petri-net.

**3.7.** Decompose density distribution of costs per activity in a phase
**3.8.** In the last step, the distribution of costs for each activity in a phase is calculated, by applying the density function over the subset of each individual activity that is in the event log of the subset.

# 4 APPLICATION IN CLINICAL CONTEXT

This chapter offers an experimental setup for evaluating the care pathways and their decompositions of patients at different stages. It contains a description of the data sources and joining of these datasets and a description of the patient population. Then it discusses the sources for costing for the different phases related to the different data sources and applies the described workflow from Chapter 3.

## 4.1 Data sources

BioGrid, a data linkage platform, provided Australian clinical registries (ACCORD and TRACC) linked to Victorian hospital administrative datasets. The hospital datasets were provided for two major hospitals in Melbourne: Western Health and Royal Melbourne Hospital. These datasets are provided to the Cancer Health Services Research Unit, for a larger study to the disparities in costs and outcomes of care provided to CRC patients. The datasets provided and the years they covered are detailed in Figure 10.

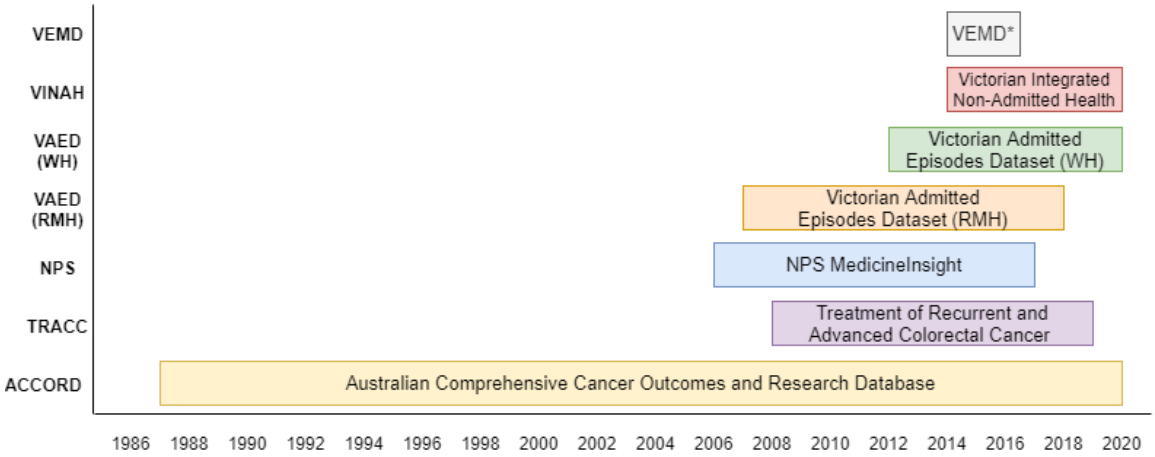



***Figure 10: Linked datasets provided by BioGrid (*Victorian Emergency Minimum Dataset)***

As the focus is on the full care pathways, we will subset the datasets to a time period for which there is data available in each of the main datasets. This is 2007-2017 for the Royal Melbourne Hospital and from 2012 to 2017 for the Western Health subset. The available Victorian Emergency Minimum Dataset (VEMD) and Victorian Integrated Non-Admitted Health (VINAH) datasets only contain a small number of years; therefore they will be excluded from discovery of the care pathway.

### 4.1.1 ACCORD

The ACCORD (Australian Comprehensive Cancer Outcomes and Research Database) collects information relating to cancer patient's diagnosis, treatment, and outcomes, for 29 tumour streams [83]. The dataset only includes patients with a primary CRC and the data will contain the basic characteristics of all the patients with CRC, as well as information on their tumour and their treatment. The dataset contains 16 tables, linkable with ID's associated to a person. Only one identifier (the `USI` or unique swap identifier) is truly unique for a patient and there are 7533 unique USI's in ACCORD. An overview of the linkage map can be found in Appendix 2, Figure 57.

### 4.1.2 TRACC

The Treatment of Recurrent and Advanced Colorectal Cancer (TRACC) dataset has enrolled a large number of patients with CRC with a metastatic or recurrent local progression, from 30 Australian and Hong Kong based hospitals, of which the Australian population is used in this study [84]. The dataset contains in-depth information on the treatment of 1422 of these patients, including clinical reasoning and medical history. The dataset contains 37 data tables, also linkable with ID's, with a hierarchical main-and-sub structure. An overview of the linkage map can be found in Appendix 2, Figure A1. The information in TRACC can be used to analyse differences in health outcomes, based on either their medical history or choices in the treatment provided. It will however be used to indicate the state of care delivery a patient is in during a period of time.

### 4.1.3 VAED

The Victorian Admitted Episodes Dataset (VAED) provides information on the use of health-services in Victoria, next to the causes, effects and nature of the associated illness [85]. This large dataset (506.605 unique patients) contains information on all patients which had encounters with the healthcare system, in which they were admitted to either a public or private hospital, an extended care facility or a day procedure centre. All data for patients admitted to the Western Health hospital (years 2012-2020) and Royal Melbourne Hospital (years 2006-2019) are available. The VAED contains a single (large) data table, and no linkage was needed.

### 4.1.4 NPS

The NPS Medicines Insight database is a primary care database, consisting of routinely taken and de-identified electronic health records (EHR's) from consenting Australian general practices [86]. During the timespan of 2006-2019, over a million (1.435.112) patients were registered in this dataset. The NPS dataset contains 17 data tables, where the patient datafile contains the basic information, including a USI in the main patient dataset. The patient datafile also contains a linkable patient ID, to the other 19 data tables, that contain for example requested investigations and prescribed medications. An overview of the data tables and linkage can be found in Figure 11.

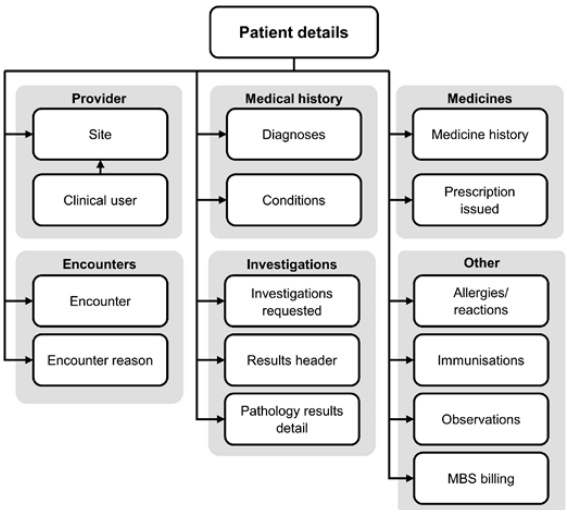



***Figure 11: Overview of data tables in NPS***

## 4.2 Patient identification and selection

A Euler diagram was produced with R's VennDiagram package, to evaluate the overlap between the patients in the different datasets. The result is displayed in Figure 13. From this overlap, a group is selected for this study.

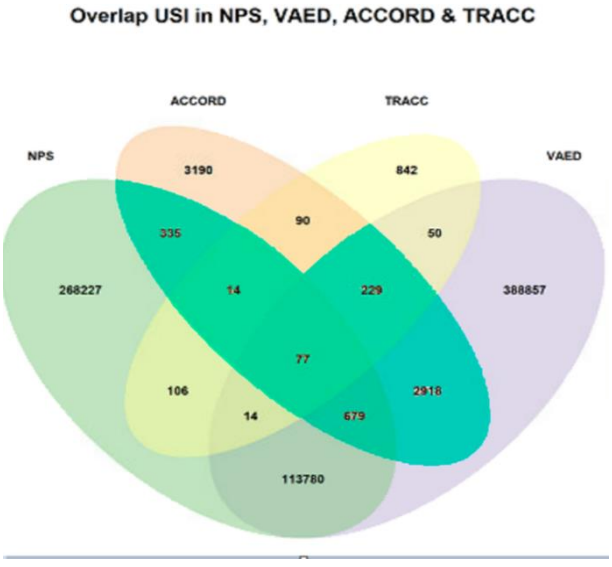



***Figure 13: Euler diagram for overlapping USI (unique patient identifiers) in the different datasets.***

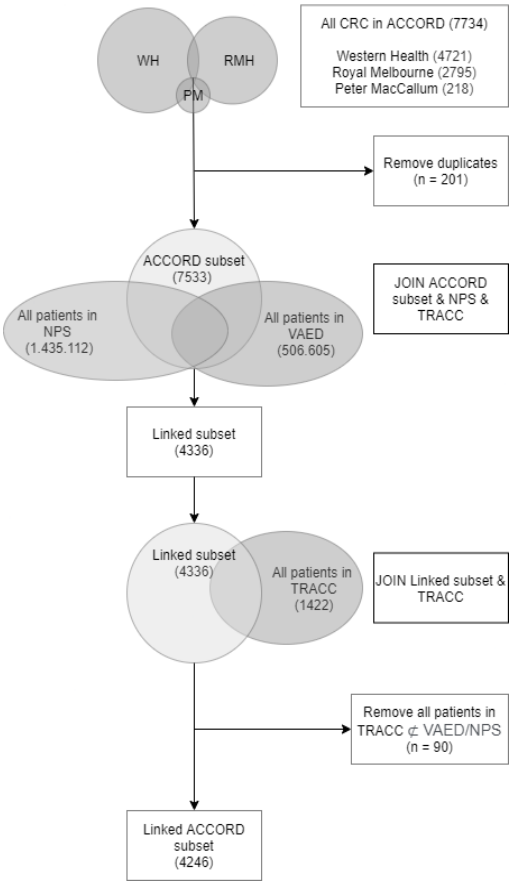


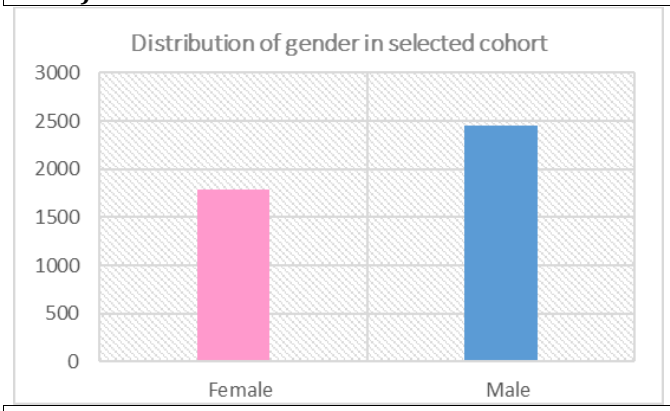

***Figure 12: Overview of the approach for patient population selection***

From the Euler diagram, a step-by-step approach is used to select patients that have sufficient information to create their pathways. A graphical overview of the steps used is provided in Figure 12. The datasets contain different patient populations and are very different in size, so we will need to link only patients that we include in our research. We use the patient USI's in the ACCORD registry as the backbone of our patient population selection and will merge these to the patient USI's in the admitted episodes in VAED and the patient USI's in the encounters with primary care in NPS. We observe that 1105 patients from ACCORD are available in NPS and 3896 patients from ACCORD are available in VAED. As there is some overlap in patients both NPS and VAED, we find 4313 CRC-patients that are available in all of these datasets. As we would like to add information on the treatment of local recurrent and metastatic cancers, we link this to all patients available in TRACC. After doing this, we exclude 90 patients that are only available in TRACC and ACCORD, that did not have any entries in NPS/VAED.

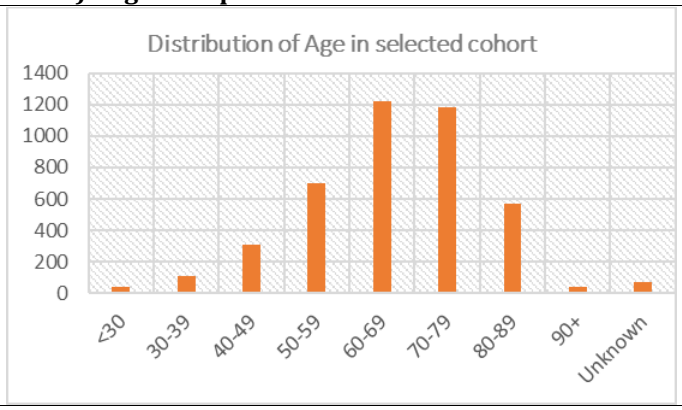
## 4.3 Population description

For the included patients, exploratory reviewing of their characteristics is presented. As we would like to identify differences in care delivered to different sub-populations, we first will evaluate baseline group characteristics, such as sex, age group, primary staging, primary tumour location, indigenous status and rurality. The distribution of the different characteristics is given in Figure 14a-f, with a breakdown by data table in Appendix C.

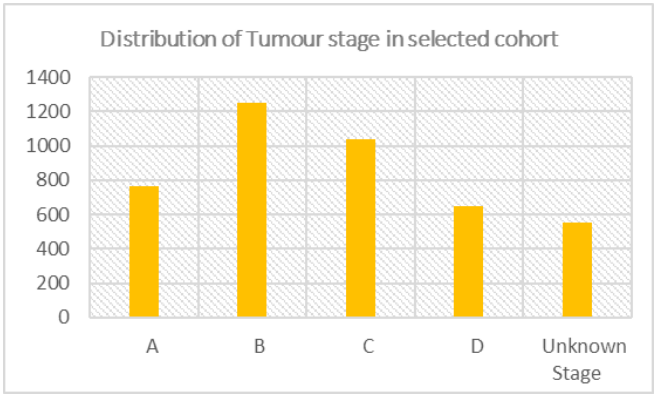
**a) Gender**



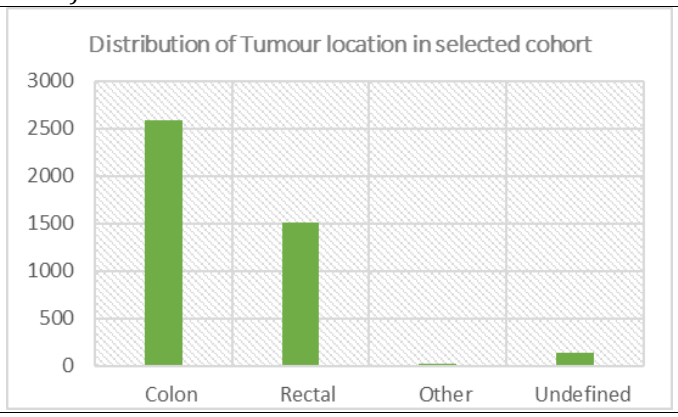

**b) Age Groups**



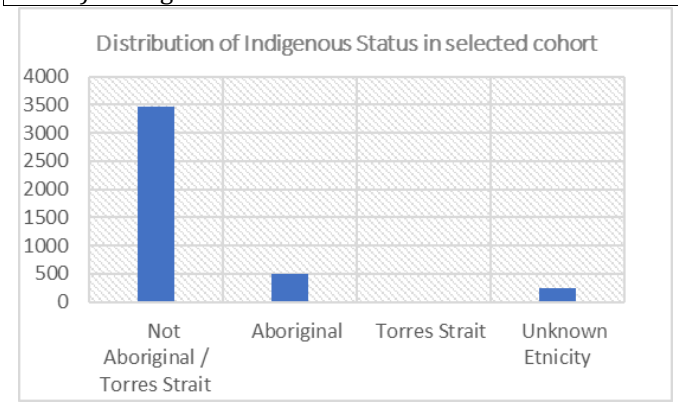

c) ACPS Staging




d) Colorectal Cancer location




e) Indigenous status




f) Remoteness

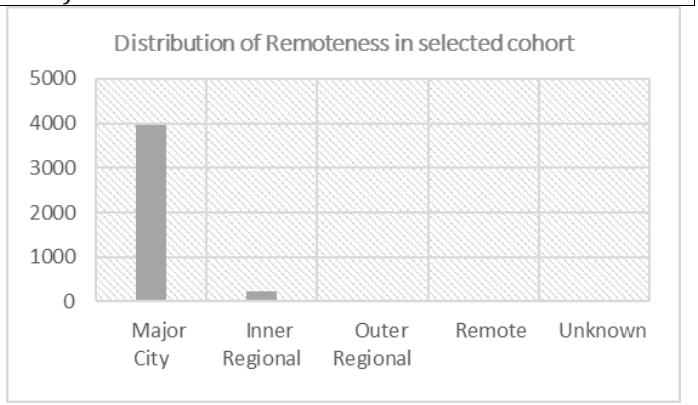



***Figure 14a-f: Patient characteristics in selected cohort***

## 4.4 Classification phases of care delivery

Within the context of care related to CRC, we identify four main phases of care delivery, that are roughly similar to the four major cost-driving phases available in the data-sources. An activity $ak(i)$, will be in one of these phases. Next to this, some milestones within a patient's life are registered as an 'Activity', and grouped as a phase, which can be used to determine in which phase of the patient's care delivery is located. A description of the five established phases is given below, with examples given in Table 6.

- The **Indication** phase contains activities that a patient receives when he/she has a cause for finding CRC. This may be because they appear symptomatic, with this patient seeking medical attention from his GP. Some patients might not be diagnosed with CRC in the first line care, but for example by screening efforts, or when the CRC was suspected within an unrelated care-episode. The inclusion of an indication element is based on the reason for encounter. The list of valid encounter reasons is based on the list of presenting symptoms and can be found in Appendix D
- The **Diagnosis** phase contains diagnostic assessments that are requested within an encounter with a Primary Care deliverer, or a medical specialist within the context of detecting and diagnosing the CRC. Expected activities are scans, lab work, histological staging and more. These activities may be part of an admitted episode and billed as such but can also be requested by a GP and be administrated in this way. The inclusion of Diagnosis elements is based on the same list as the list of reasons for encounter.
- The **Admitted Episodes** phase contains hospitalisations. Within the care pathway of most cancer patients, they are admitted into a hospital or equivalent centre at least once. In Victoria, admissions registered in the VAED can be both for day-treatment as well as overnight stays. Within these admitted episodes, various diagnostic tests, medications, and procedures may be provided to a patient. The inclusion of an admitted episode is based on its DRG (Diagnosis Related Group) description, based clinical expert opinion.
- The **Medication** phase contains medication that can be prescribed by both primary care givers for regular usage, and specialists in a hospital setting as part of an admitted episode. This also includes the use of drugs for chemotherapy, which are recorded separately in the ACCORD registry.

***Table 6: Type of activities in a phase of a process within a care pathway context***

| Phase | Description | Examples |
|---|---|---|
| Indication | Reasons for Encounter | • GP-visit<br>• Referral from screening<br>• Detected in another admission |
| Diagnosis | Requested diagnostic tests for a CRC-diagnosis | • Imaging Techniques (CT/MRI)<br>• Colonoscopy<br>• Histology |
| Admitted Episodes | Hospital admissions with a relation to CRC | • Surgeries<br>• Admission for Chemotherapy<br>• Palliative care |
| Medication | Prescribed medication related to CRC treatment | • During Chemotherapy admission<br>• Prescribed for symptoms |
| Life-Events | Life Events related to CRC | • Diagnosis<br>• Death<br>• Lost to Follow-Up<br>• Survivorship |

### 4.4.1 Element derivation per phase

In the different phases and associated datasets, an activity we are interested in, can be formatted differently in the health records. Some elements, such as deaths or diagnosis are registered only by a variable in date-type. Other activities, such as chemotherapy regimens, may constitute of multiple cycles of administrating a combination of drugs, all with distinct dates. In Table 7, an overview is given of how each type of activity in each dataset is derived.

For the Indication phase, the MBS Item description from the MBS dictionary was linked as the name for that specific encounter. For diagnosis, the free text field was separated by separation tokens, and then the resulting names were coupled to an item number. In the admitted episodes dataset, the Hospital DRG was linked to a dictionary of DRG descriptions. For the medication in the NPS prescription dataset, the names registered as the medicine name were used. For the chemotherapy in ACCORD, the individual regimen names were used for every cycle. Then, this regimen was linked to every prescribed medicine in that cycle and to the associated episode, such that cost could be aggregated. For the life events in both ACCORD and TRACC, the column containing the date of death was used to indicate a passing and the patients' status was transcribed on the last encounter date as an activity.

*Table 7: Overview of Attributes in datasets for 5 phases and two events and resulting traces*

| Phase | Dataset | Attributes of interest | Resulting Trace (Example) | | |
|---|---|---|---|---|---|
| | | | **Case-ID** | **(Event-ID) EventName** | **Timestamp** |
| Indication | NPS_encounters | Encounterdate<br>Encounter_Reason<br>Item Number<br>MBS Item Description | 00001 | (10990)<br>Management of Bulk-Billed Services | 2008-02-01 20:00 |
| Diagnosis | NPS_Tests | Request_Date<br>Encounter_Reason<br>Item Number<br>Requested_Tests | 00001 | (10990)<br>FBE + LFT + FATS + Glucose | 2009-03-01 12:00 |
| Admitted Episodes | VAED | Admissiondate<br>Dischargedate<br>Hospital DRG description<br>WIES_Value | 00001 | (GO1B)<br>RECTAL RESECTION - CCC | 2015-03-01 12:00 |
| Medication (GP) | NPS_Prescriptions | First_date<br>Encounter_Reason<br>MD_No<br>Medicine_Name | 00001 | (-)<br>Dexamethasone | 2015-03-01 12:00 |
| Medication (Chemotherapy) | ACCORD | Chemo_Episode_Start_Date<br>Chemo_Episode_End_Date<br>Chemo_Cycle_Start_Date<br>Chemo_Cycle_End_Date<br>Cycle_Number<br>Total_Cycle<br>Regimenname<br>MedicationID | 00001 | (CycleN)<br>FOLFOX | 2015-03-01 12:00 |
| Life-Events | ACCORD | DateOfDeath<br>Last_encounter<br>Status | 00001 | Death | 2015-03-01 12:00 |
| Life-Events | TRACC | DateOfDeath<br>Status<br>Last_encounter | 00001 | Disease Progression | 2015-03-01 12:00 |

## 4.5 Determining costs for care activities

Within the four main phases of care delivery, different sources were used for costing the elements in that phase. Two main sources of information were used: Information captured in the NPS dataset could be linked to the Medicare Benefits Schedule (MBS) and Pharmaceutical Benefit Scheme (PBS), which register prices for medical services and medicines [87], [88]. For hospital care in the VAED, a value is registered that is related to an entire admission, by the Weighted Inlier Equivalent Separation (WIES) methodology of costing.

### 4.5.1 Medicare Benefits Schedule (MBS) and Pharmaceutical Benefit Scheme (PBS)

All medical services and medicines captured in the NPS MedicineInsight dataset, were priced according to the corresponding item numbers and drug codes as listed in the Medicare Benefits Schedule (MBS) and Pharmaceutical Benefit Scheme (PBS) respectively. Both the MBS and the PBS provide online tools, in which item numbers and drug numbers can be found. Web scraping with the RSelenium package was used to retrieve the costs of the standard units of MBS items and PBS medication numbers. As item numbers and codes may change over time, iterative checks were conducted to ensure all codes were adequately captured. Items and medications that comprised less than 1% of all registered item numbers (<0.05%) and prescriptions (0.8%) were excluded as they could not be adequately identified. The NPS prescription data table contains not only the specific drug name, but also the number of repeats, the dosage and the container size. From the PBS webpage, the standard unit costs, the maximum allowed prescription amount were obtained and multiplied with the prescribed dosages and repeats. In this manner, the costs for each prescription was calculated. The prices of both the MBS item numbers as well as the PBS prescription numbers are based on the latest year of registry (2020).

### 4.5.2 Weighted Inlier Equivalent Separation (WIES)

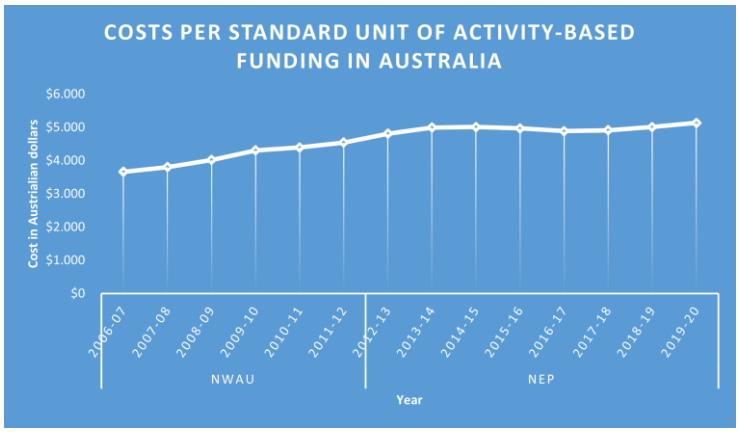



***Figure 15: Value of a standard unit of delivered care in Australia, before 2012 called the National Weighted Activity Unit (NWAU) and after 2011 the National Efficient Price (NEP)***

The Australian healthcare system, like some European healthcare systems do, makes use of activity-based funding, specifically related to diagnostic related groups, or DRG's for billing care-activities within an episode of admission. The current used system for the included hospitals is called the WIES and was implemented in 2011. This system is related to the Australian National Efficient Price (NEP), which is a nationally set price for a unit of care, and is adjusted every year by the Independent Hospital Pricing Authority (IHPA) [89]. Before the introduction of the current system, the National Weighted Activity Unit (NWAU) was used, which had a slightly different method of calculation. The Victorian Department of Health's current funding model is called WIES, which stands for Weighted Inlier Equivalent Separation, and is used to calculate a factor, which represents the fraction of a single care-unit. This fraction can be multiplied by the value of a single WIES, which is roughly equivalent to the NEP, to obtain the costs of that admitted episode. The extensive WIES calculation places the patient in a Diagnostic Related Group, or Major Diagnostic Category, based on body system and principal diagnosis. Further sub-classification are based on the presence of complications or comorbidities and age. Each DRG has an `acceptable` range of Length of Stay. If a patient is discharged within this timeframe (called an Inlier), a fixed amount is paid. WIES payments for low outliers are discounted and high outliers receive additional WIES for each day that their length of stay exceeds the high boundary point of inliers. This results in a factor, which is registered in VAED. In the calculation of the value of a DRG in the VAED, the factor registered is multiplied by the value of the standard unit of activity-based funding, the NWAU value prior to 2011 and the NEP value (see Figure 15). All costs are reported in 2020, and adjusted with the average Consumer Price index for Health from the Australian Bureau of Statistics in that year [90]. The CPI indices as well as the value for NWAU and NEP can be found in Appendix 5.

### 4.5.3 Application to phases

The previously defined phases contain elements where specific costs need to be added to. The source of these prices varies per phase and is related to which information in captures in the corresponding dataset.

The elements in the **Indication** phase have costs incurred when a patient receives an indication for searching for CRC, by visiting their GP. Other costs might have been incurred when CRC was not found in the first line care, by screening efforts, or when the CRC was suspected within an unrelated care-episode, but costs from screening or related to other care-episodes are not registered. For costing the GP-encounters, the registration of the item number in MBS, or Medicare Benefits Schedule and its associated price are used.

The **Diagnosis** phase contains scans, lab work, histological staging and more. These activities can also be part of an admitted episode and billed as such, but they would be included in the bulk-costs of an admitted episode. They can also be requested by a GP and be administrated in this way. Costs for diagnostic tests requested on a GP visit are included, by separating the free text field "Requested Tests" in the prescriptions table to all individual elements. For example, 'FBE + LFT + FATS + Glucose' contains 4 elements within one Diagnosis request. These individual elements are coupled to specific item numbers in MBS and its associated costs aggregated over all elements.

The **Admitted Episodes** phase contains hospitalisations with a unique DRG, with an associated WIES-fraction. E.g: An instance of DRG 'Complex Gastroscopy' can have a WIES factor of 1.345 in 2015, where the value of a WIES is $5.007 and the CPI = 5.5%, the costs of this DRG is: 1.345 * $5.007 * 1.055 = $7104.81 For the admitted episodes, there will not be a breakdown in costs over the many registered procedures in a DRG, as there the WIES costing methodology does not have standard values for procedures.

The **Medication** phase contains medication that can be prescribed by both primary care givers for regular usage, and specialists in a hospital setting as part of an admitted episode. While in a primary care setting (NPS) the prescribed drugs are registered separately, in a hospital setting, the medication administered will be booked as part of the DRG and it will not be in the records. This is not the case for chemotherapy drugs, which are recorded separately and can be obtained in the ACCORD data file.

***Table 8: Costing methods for five distinct phases in a care pathway***

| Phase | Dataset (table) | Example elements | Costing Source | Based on |
|---|---|---|---|---|
| Indication | NPS (Encounters) | GP-visit | MBS | **3.9.** MBS Item number for GP encounters<br>**3.10.** MBS Fee per registered Item number |
| Diagnosis | NPS (Requested Tests) | Imaging Techniques (CT/MRI)<br>Colonoscopy<br>Histology | MBS | **3.11.** MBS Item number for requested tests<br>**3.12.** MBS Fee per registered Item number |
| Admitted Episodes | VAED | Surgeries<br>Admission<br>Chemotherapy<br>Palliative care | WIES | **3.13.** WIES Factor<br>**3.14.** WIES value of year of administration<br>**3.15.** CPI of year of administration |
| Medication (GP) | NPS (Prescriptions) | Prescribed for symptom | PBS | **3.16.** PBS Drug code / name for prescribed medication<br>**3.17.** Dosage<br>**3.18.** Repeats<br>**3.19.** PBS Fee per maximum dosage |
| Medication (Chemotherapy) | ACCORD | Prescribed for Cycle of therapy | PBS | **3.20.** PBS Drug code / name for Chemotherapy Medication<br>**3.21.** Dosage<br>**3.22.** Repeats<br>**3.23.** PBS Fee per maximum dosage |
| Life-Events | TRACC/ ACCORD | Diagnosis<br>Death<br>Lost to Follow-Up<br>Survivorship | No costing | |

# 5 RESULTS

The aim of this chapter is to provide a comprehensive analysis of the process mining methodology applied to colorectal cancer pathways. For the analysis, a supplementary output document is created, which can be used interactively. With this app, panning and zooming in on the resulting images is possible, which for some of the graphs is essential. At the time of writing, the app is available on: https://strelijveldstudent.shinyapps.io/R_Outputfile/). In Chapter 6, we will address a specific case study, evaluating the differences between the pathways of colon cancer patients in the different ACPS stages.

In the first part of the results, Section 5.1, we evaluate the resulting pathways and describe their Petri net components in Section 5.1.1. In section 5.1.2 we evaluate the enhanced pathways, displayed as direct follow graphs and in section 5.1.3 we study the validity of the observed pathways for the different subpopulations in a quantitative manner, using the conformance metrics. With the results of 5.1.1 and 5.1.2 we are able to answer the first research question '*How can a care pathway for CRC patients be derived with process mining techniques and how can these care pathways be evaluated?'*. Section 5.1.2 then provides an answer to the second research question *'What are the characteristics of the patients going through the (main) CRC pathways?'*. In the second part of the results, Section 5.2, we evaluate the cost-enhanced pathways obtained with the custom algorithm from Section 3.2, as well as cost-enhanced direct-follow graphs and answer the last research question: *'How can the total costs of the (main) CRC care pathways be calculated and evaluated from a cost-perspective?'*.

## 5.1 Comparative analysis of complete care pathways

Care provided to patients will be evaluated with resulting Process models for the derived pathways. The pathways are evaluated in each of the four main phases (Admitted Episodes), Medication (Chemo Episodes & GP Prescriptions), Diagnosis (Requested Tests), and Indication (GP visits). Resulting calculation times are provided in Appendix 6. In this section, one of the resulting Petri nets of the care pathways is displayed and different parts are highlighted and explained, so that the readers can familiarize themselves with the visualizations and how to interpret them.

### 5.1.1 Introduction and visual inspection of resulting care pathways

The complete pathways presenting the five different phases, based on the entire cohort of patients is displayed in Figure 16 (Admitted episodes), Figure 22 (Chemo Episodes), Figure 23 (GP Visits) and Figure 24 (Diagnostics and Prescriptions). The generated images can be too large to read, and it is therefore required to zoom in to view the textboxes containing the activities. Figure 16, displaying the pathway of only the care activities in admitted episodes will be used as an example to describe the different model components.

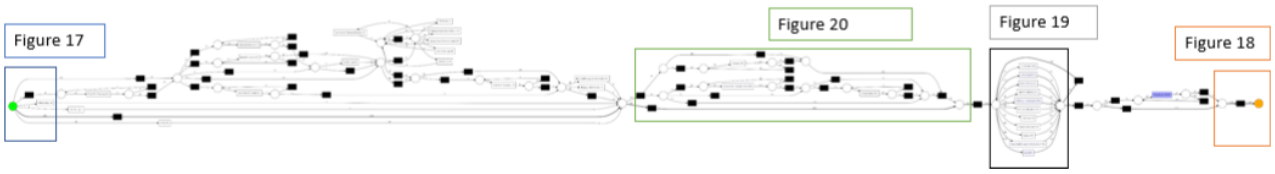



***Figure 16: Resulting pathway of phase: Admitted Episodes.***

#### 5.1.1.1 Explaining Model Components: Legends to the figures

First, we evaluate some of the structures and annotations in the model that are necessary to interpret these models correctly. The first components of interest in the model are located at the start and end of the pathway. The green circle, indicating the starting place (see Figure 17 on the next page) and the orange circle, indicating the end place (see Figure 18 on the next page).
The numbers displayed next to the arrows, represent the number of possible unique pathways that passed through the previous location and continue in the direction of the arrow. Multiple patients can have the same pathway, and if this is the case, this number does not increase.

We see that a total of 3098 unique pathways are registered at the end of the model. In the pathway, we see transitions (the boxes), containing the name of the activity. For example the care provided for DRG 'GI Haemorrhage' in Figure 17, which has a number, in this case '4' after this. This number represents the total number of times this activity was performed, and it does increase with every patient and can even increase multiple times from a single patient. The black rectangles are the 'silent' transitions, which do not represent an activity, but are necessary to display loops, skipped steps or simply no activity until the pathways converge again.

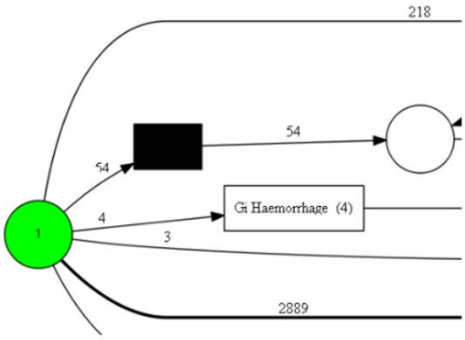



***Figure 17: Beginning of the care pathway of admitted episodes***

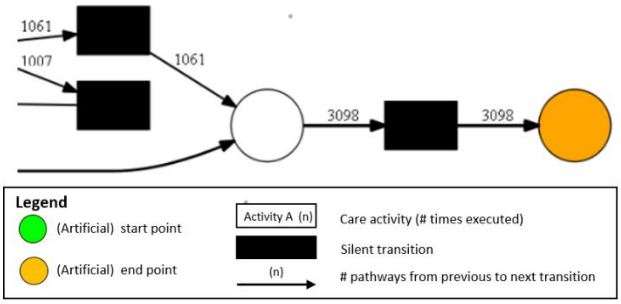



***Figure 18: End of the care pathway of admitted episodes***

In Figure 19, we see a part in the model where concurrent behavior with a loop is displayed. The coloring of the transition is based on the number of times this activity is performed, relative to the total number of all activities performed in the net. The larger this number is, the darker the shade of blue. From the starting point there are 11 possible activities or scenarios that can happen at this point in the model, each with their individual number of pathway versions that go through this activity. Importantly, these activities are not mutually exclusive. A patient can first have one of the procedures belonging to a DRG such as the 'Major small & Large bowel procedure' and then have a follow up, perhaps with the addition of a Colonoscopy. In the model, we see that a silent transition is added, which is connected to the starting place ( the red line), which secures that it is possible to return to this place.

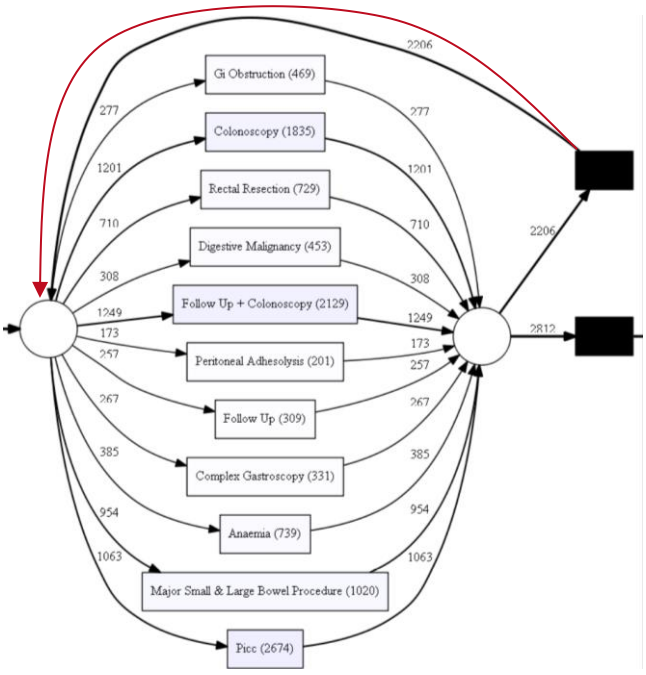



***Figure 19: Concurrent parts in the pathway***

In Figure 20, we see an example of specific complex behavior of the model. After the place where all pathways converge (the blank point left in the model), four scenarios are possible, including a 'no-activities at all'-scenario. As the number of pathways starting from this point are 258 and 2051 respectively, patients do not have the same pathway in the parts before and after the converged point at the left. For the three activities in this complex part, Colostomy, Abdominal Pain Mesentrc Adents and Gi Haemorrhage, we see that in any case, the Gi Haemorrhage-procedure is performed later than the other two, and that the earlier two are mutually exclusive. We also see that all three of the DRG's have their individual loop, such that it is possible that patients would have the procedures belonging to that DRG more than once. Lastly, we see that all of these DRG's are possible, but not essential in the pathway, as they have silent transitions concurrent to the activity (see green line), and to all the activities together (blue line).

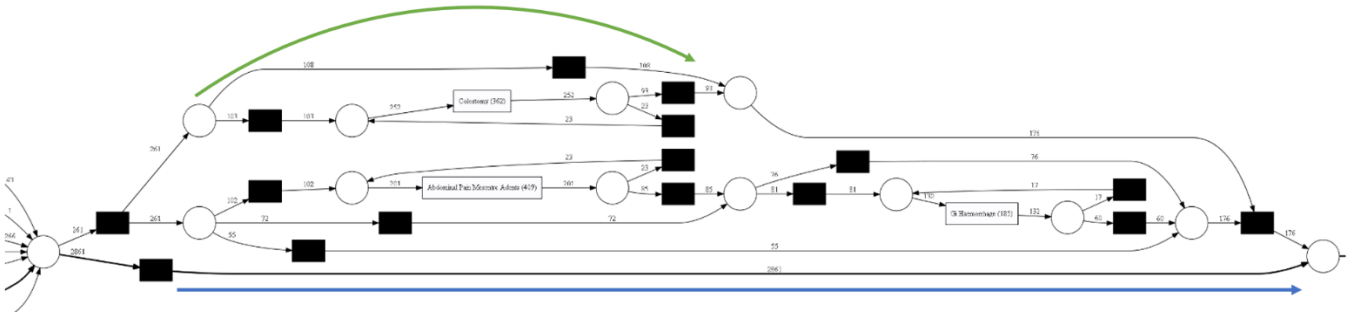


***Figure 20: Example of complex model behavior***

### 5.1.1.2 Model Evaluation

Within the resulting models, we can see different parts where pathways converge and diverge. Within the pathway of a stage of care, such as the Admitted Episodes, we can therefore distinguish different sub-pathways within the entire model. In **Error! Reference source not found.** we see 8 of those, with a label (**A through H**) attached to it.

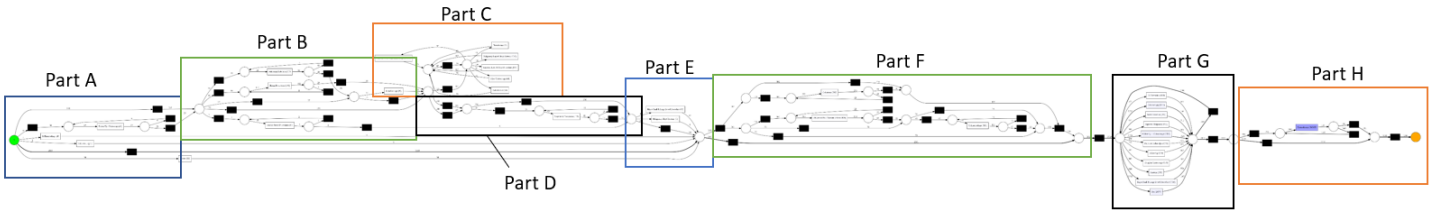



***Figure 21: Parts in pathway Admitted Episodes.***

In **part A**, we see that pathways are possible that include 'Gi Haemorrhage', a 'Follow-up with an endoscopy', perhaps multiple times, and 'Unknown DRG'. We also see that in 'GI Haemorrhage' and the unknown DRG, the number of pathways through the activity is equal to the number of times the activity was performed, so no repeats and each of the pathways through the activity is unique. With the 'Follow Up + Endoscopy', we see that 63 different pathways go through the activity and six a second time via the loop, yielding a total of 69 activities of 'Follow Up + Endoscopy' performed. **Part B** is a part where procedures are performed for 'Pulmonary Embolisms', 'Hernia procedures' and 'Anal & Stomal procedures', with the possibility to follow the Pulmonary embolism procedure with a Bronchoscopy. The three DRG's all have loops, suggesting that there can be multiple admitted episodes for the same DRG. **Part C** is somewhat of a sidestep, with multiple possible procedures that are cancer related, such as Radiotherapy, bone and liver malignancies or related procedures, or thrombosis because of colorectal cancer. **Part D** only consists of a possible DRG 'Respiratory neoplasm', which might take place after the previous procedures. **In Part E**, a single DRG for a Hepatic systemic malignancy, as well as a Minor Small & Large bowel procedure are found, without repeats. After these, all pathways converge and expand into **Part F**, where we observe the main DRG's for surgery performed for CRC, the Colostomy and the 'Abdominal Pain Mesenteric Adenitis', which is inflammation of the lymph nodes in the mesenteric region, likely due to the cancer. After the latter one, a procedure to contain GI Haemorrhage can be performed as well. All these procedures can be repeated, again suggesting that multiple admissions might be necessary to fulfill the objective.
In **Part G** we see concurrent options, of which most are related to either bowel imaging or follow up (Follow Up, Follow Up + Colonoscopy and Complex Gastroscopy) or to a surgical intervention in the bowel region (Rectal Resection, GI obstruction, Peritoneal Adhesolysis, Digestive Malignancy, Major Small & Large Bowel Procedure). An exception is the resulting Aneamia-related procedure and the PICC (peripherally inserted central catheter) used for chemotherapy.

**Part H** is solely consisting of admission for administrating Chemotherapy, possibly multiple times, which is also by far the most frequently occurring activity in this pathway.

The pathway of the Chemo Episodes (Figure 22) phase has a less complex structure and contains only 3 types of elements. There is a start of the chemotherapy regimen (**Part A**) after which the loop is entered and followed by one of the chemo regimens (**Part B**). In the latter of those, administration of the regimen can be repeated with a silent transition. We observe that the algorithm has placed the 'Start Chemo Episode' activity not in front of the concurrent loop, while this would be expected, as all these should be timed before administration of the regimen. This might be explained by the fact that the timestamp of starting the next line of chemo coincides with the end date of the previous and the algorithm picks the most frequent first. There are a few activities that only happen once, without a loop (see **Parts C**), after which the concurrent section ends. We observe that the resulting pathway does not yield a specific order between the 1st, 2nd or 3rd – line treatments, while this would be expected

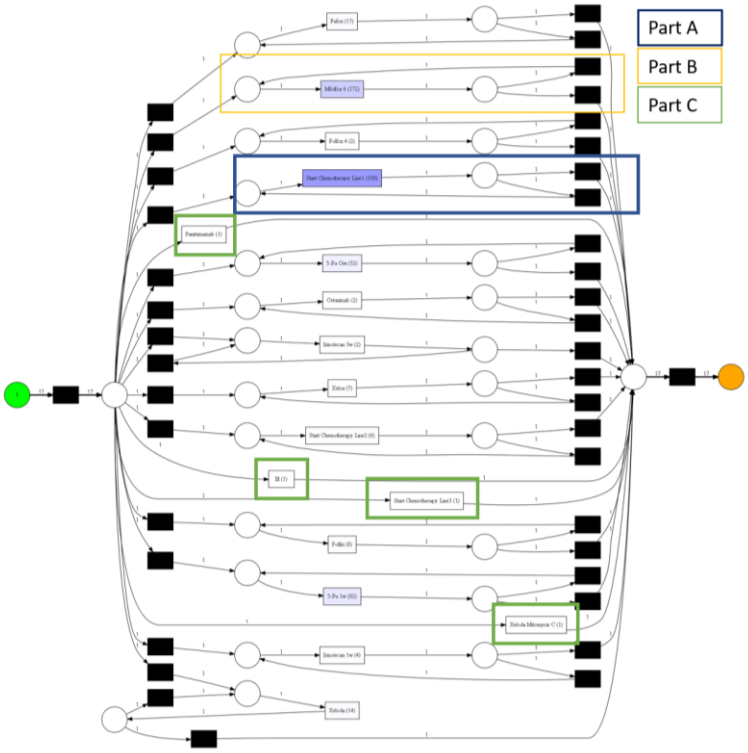



***Figure 22: Parts in pathway Chemo Episodes***

In comparison, the GP visits-pathway in Figure 23 does have a specific order and consists of three sub-pathways. Part A is unique, only 1 patient had this component, which is mental health attendance. Part B consists of 'Special' types of GP attendance, such as After-Hours attendance, Health assessments, multidisciplinary care, or services provided by either a practice nurse or a health practitioner of Aboriginal or Torres Strait Islanders descent. These variations are rare compared to the standard practices in part C, where we find either the GP's attendance or the registry of Bulk-billed services, which is a fee or reduction of costs for when multiple services such as imaging or diagnostics are declared at once. We see that these can both happen multiple times and they are mutually exclusive.


***Figure 23: Parts in pathway GP Visits***

The Prescription and Diagnosis pathways, both from the NPS data (see Figure 24) both have a similar layout and only contain two unique elements; a single activity (example in **Part A** and **Part C**) and an activity that is repeated through a loop (example in **Part B** and **Part D**). When zooming in on the individual activity, we see that most all have an occurrence for a single patient and a handful have a low number of occurrences for an equally low number of patients. With a large number of unique activities for unique cases, an order or sequence cannot be found.

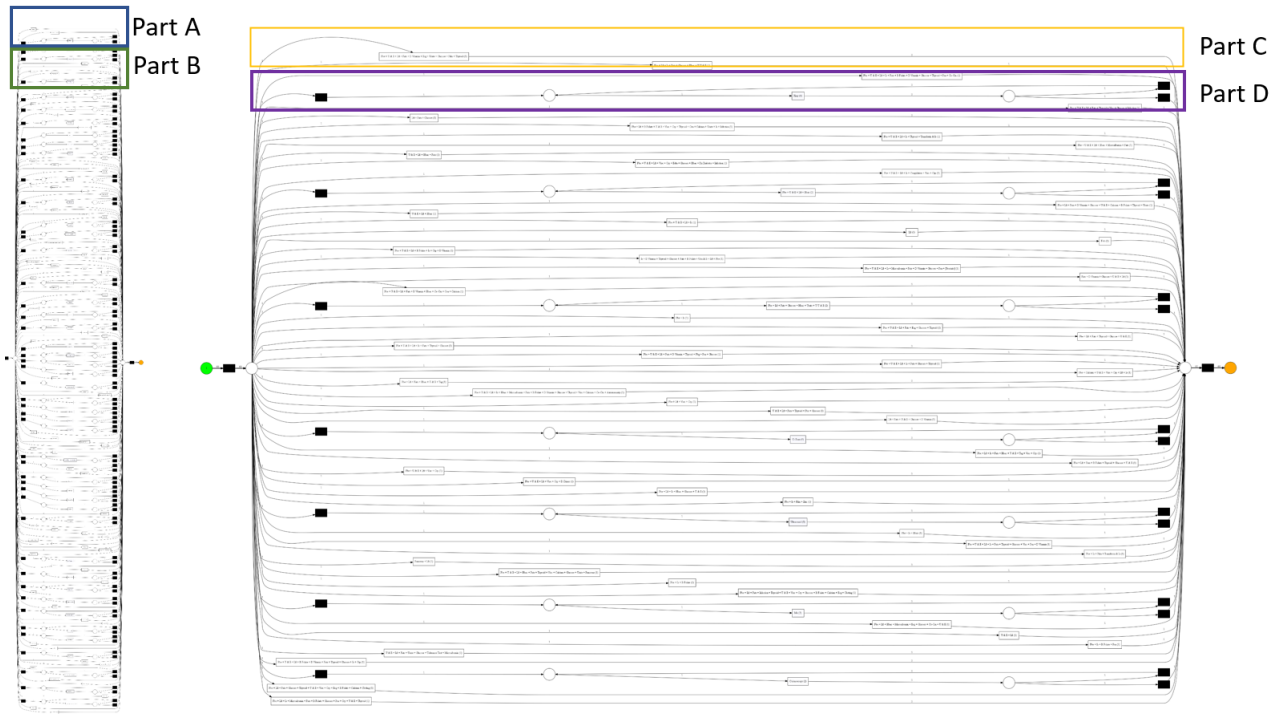



***Figure 24: Parts in pathway Prescriptions (left) and Diagnostics (right)***

### 5.1.1.3 Model Alignments

Next to the different resulting models with the frequency annotation, we have generated graphs displaying alignments of a model to different subpopulations. An example is given in Figure 25. We see the same location as in Figure 19, but the annotation is now the alignment of subpopulation 'Males'. As described in Section 2.2.3, the two numbers following the activity are the number of fitted activities (synchronous moves) and the number of moves on the model that were necessary at this transition. The more moves were necessary relative to the total, the redder the color of the transition. This gives an indication how many of the pathways could not be constructed with the care activities of the 'Males' subpopulation. This is interpreted as 'how much deviance there is between the model and the group at that point in the model'. These values get aggregated to the model's total quality metrics and if models with lower quality metrics are found, then the graph with these alignments as annotation will provide the location within the pathway that is responsible for this lower score.

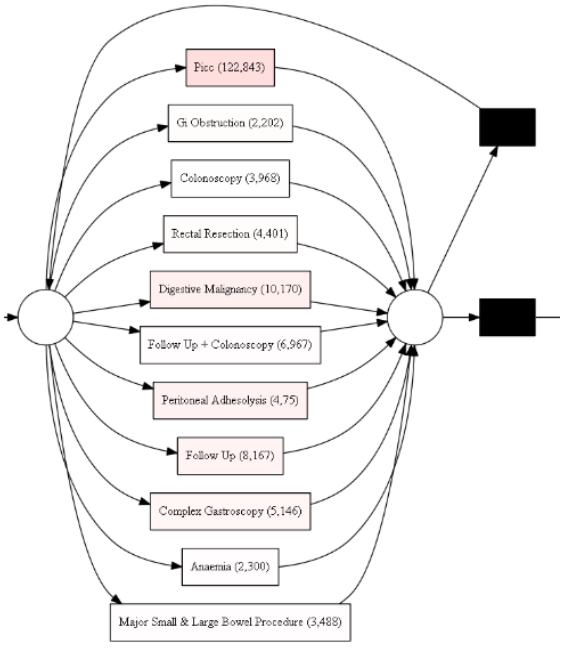



***Figure 25: Alignments on resulting model, (comparator group is all Males)***

### 5.1.2 Process Enhancement on the care pathways

In the resulting app, we can create animated graphs, using the ProcessAnimateR package from the bupaR suite, which can enhance the processes interpretation and extend it with additional information. This type of model enhancement is performed on direct-follow graphs (DFG), and not on the previously seen Petri nets. These DFG do not have the same mathematical properties as the previous models and cannot be used to align traces and for this reason are not suited to do conformance checking. However, they can be used for process enhancement. In the app, we can create these DFGs for each of the subpopulations in tab 'Basic Process maps' and use the enhanced version with animations in tab 'Animated maps' where the interactive feature to see individual patients moves through the pathway. We can display the characteristics of the patients as annotation in this animated DFG as is done in Figure 26, with the ACPS staging annotated on the pathway for GP visits. The timeline can be adjusted to an absolute value, animating on a timeline from the first to the last timestamp, or be set to relative, in which case all patients are aligned to the timestamp of their first activity. In the nodes in the graph, the name of the activity, the mean costs, and the number of times this activity was performed is displayed. On the arcs we see the percentage of patients that crossed this arc from the previous node, as well as the mean duration between the previous node and the next one.

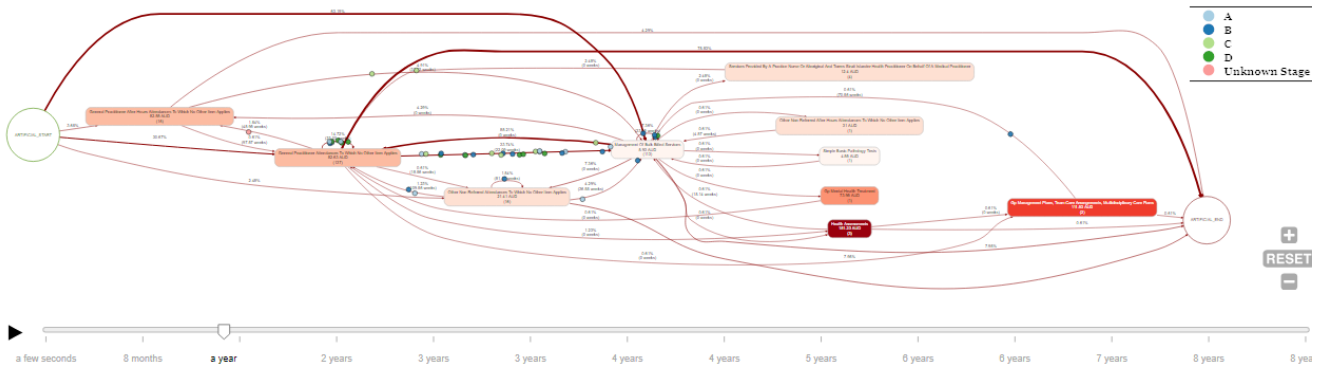



***Figure 26: Animated direct follow graph of GP Visits, annotated the with the ACPS stage the patient is in.***

In a similar fashion, another form of Process Enhancement is possible, based on additional information from a separate but related event-log. For example, the life events retrieved can be used for this. In Figure 27, the DFG of the Chemo Episodes is annotated with another event log, containing the Chemo line number that the matched patients in TRACC have at a certain point in time (named TRACC_CHEMO_LINES). We observe that the different chemo lines are visible in the graph as color annotation on each patient, with the patients that were not matched in the TRACC dataset appearing white.

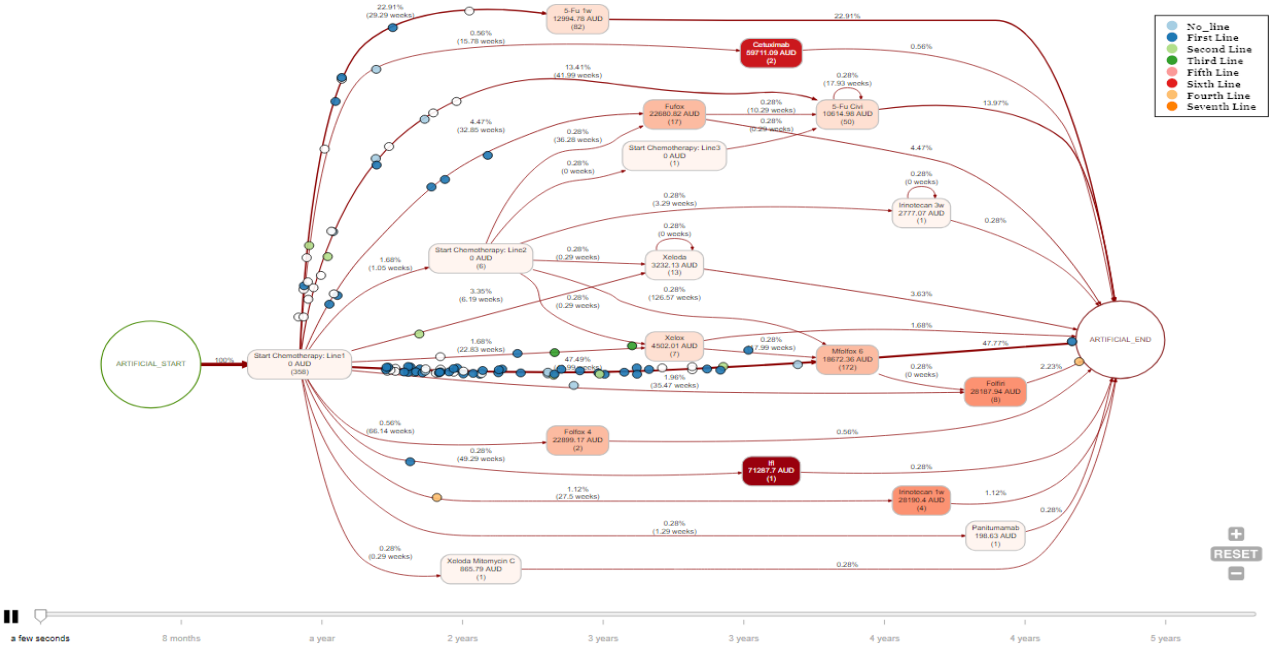



***Figure 27: Direct follow graph of Chemo Episodes, annotated with external data containing the chemo line the patient is in.***

Another possible extension is with the life events from the TRACC and ACCORD registry. TRACC contains the date of death, date of entering palliative care, the end of a round of chemotherapy, surgeries and resulting cancer outcomes, such as disease progression and patient requests to stop treatment. ACCORD contains the date of Diagnosis, the date of death, the last date of follow-up or the last registered date of care, where the patient is still alive. Also, the specific chemo regimen in TRACC is possible (see Figure 28), however, we observe that the number of unique regimens is larger than the number of colors in the pallet, resulting in double usage of colors, which results in being unable to differentiate well.

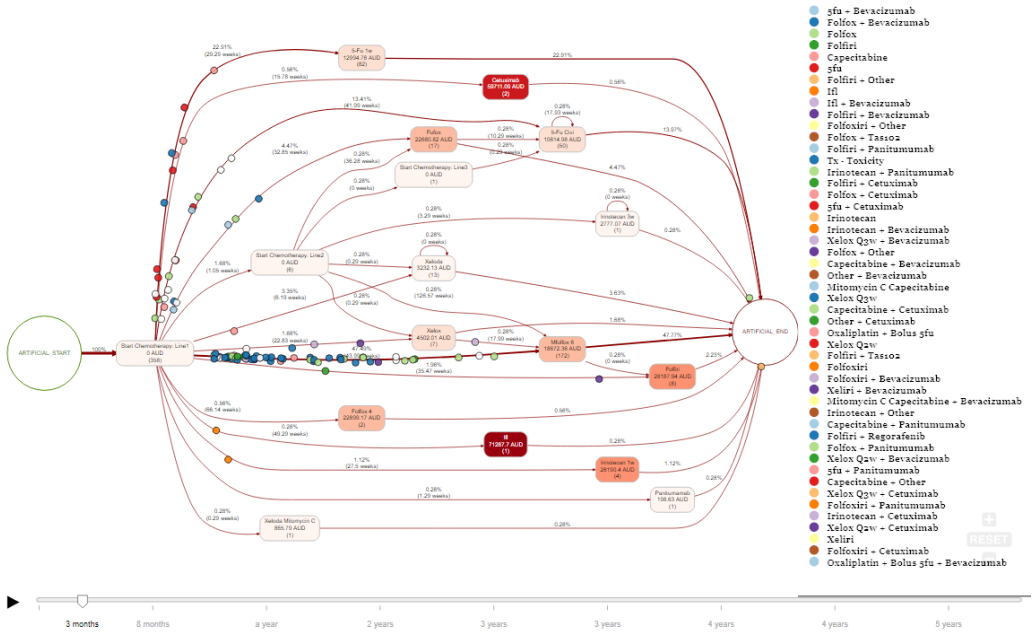



***Figure 28: Direct follow graph of Chemo Episodes, annotated with external data containing the specific type of chemo regimen that the patient is in.***

### 5.1.2.1 Conclusion first research question

Returning to the first research question:

> *How can data-driven models for care pathways for CRC patients be derived with process mining techniques and how can these care pathways be evaluated?*

We have described a workflow to preprocess patient data towards event logs, using linked data from multiple colorectal cancer registries. Then we derive process models from the care pathways, for the main phases in the colorectal cancer care pathway. This workflow contained process mining techniques for discovery, using the inductive miner algorithm to derive the process models, that could be visually inspected and interpreted.

Additionally, from the event logs Direct Follow Graphs were constructed, which could function as an enhancement of the visualizations, displaying characteristics on a patient-level. As well, these graphs can be extended with external event logs, containing information of patients that change over time, so they, for example, provide context on the patient's status or the patients treatment line or regimen.

The equivalence of pathways can partially be evaluated by comparing the resulting alignments of the models visually and this can be extended by quantitative analysis using the conformance metrics, resulting from the conformance algorithms in the workflow. A combination of both is needed to adequately evaluate the pathways that result from the process mining workflow.

### 5.1.3 Quantitative analysis of the validity of the care pathways

After we have derived process models, we would like to know how good these models explain the sequences of care activities that patients have received. Next to a face-value evaluation of the models, we can quantitatively evaluate the models using the conformance algorithms. We will research which subpopulations have their care best explained by the derived CRC pathway models, to answer the question: *What are the characteristics of the patients going through the (main) CRC pathways?* The quality of a model can be described by the four conformance metrics, *fitness, precision, generalization & simplicity*. These values, except for the simplicity, are calculated in comparison to a reference dataset, so that they explain how well the model represent the data provided in the reference dataset. Additionally, the average of these values, as well as an F-score, a balanced score for the fitness and precision is calculated. The resulting graphs are displayed in the 'quality metrics' tab of the app. First, the average conformance is of interest, provided in graphs such as Figure 29. As all resulting graphs are approximately the same, we will evaluate one dataset only, the one for chemo therapy episodes.

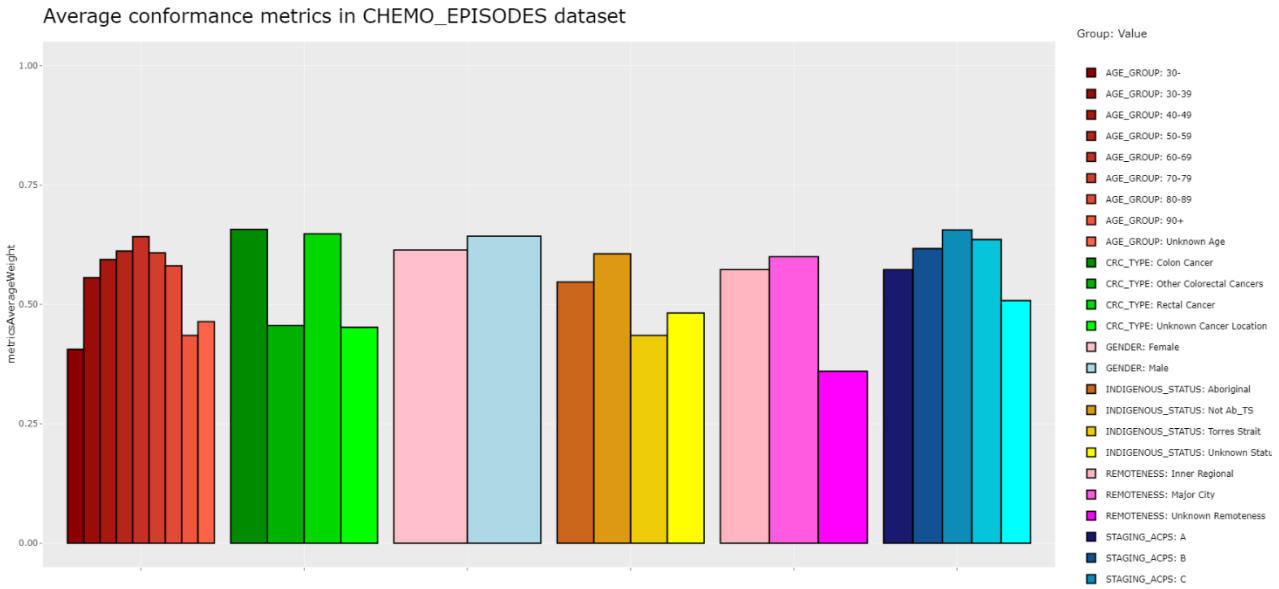



***Figure 29: Average conformance metrics of the main derived model from Chemo Episodes in comparison to each of the subpopulations***

Overall, the quality of the model is represented best in the average of the conformance metrics. The average value will be in the (0,1) interval and the population with the highest value will have a care pathway that is best represented by the derived model. This chart is used to find outlier subpopulations, for which the main model is not representative. The highest value found, 0.656, corresponds to the ACPS stage: C, indicating that patients with ACPS stage C are best represented by the model and this value is also the maximum value for the current model. The other subpopulations should then be compared with this value, so a low conforming population, such as those with an Unknown Remoteness status, with an average score of 0.360, is approximately half as well represented by the model as the Stage C patients.

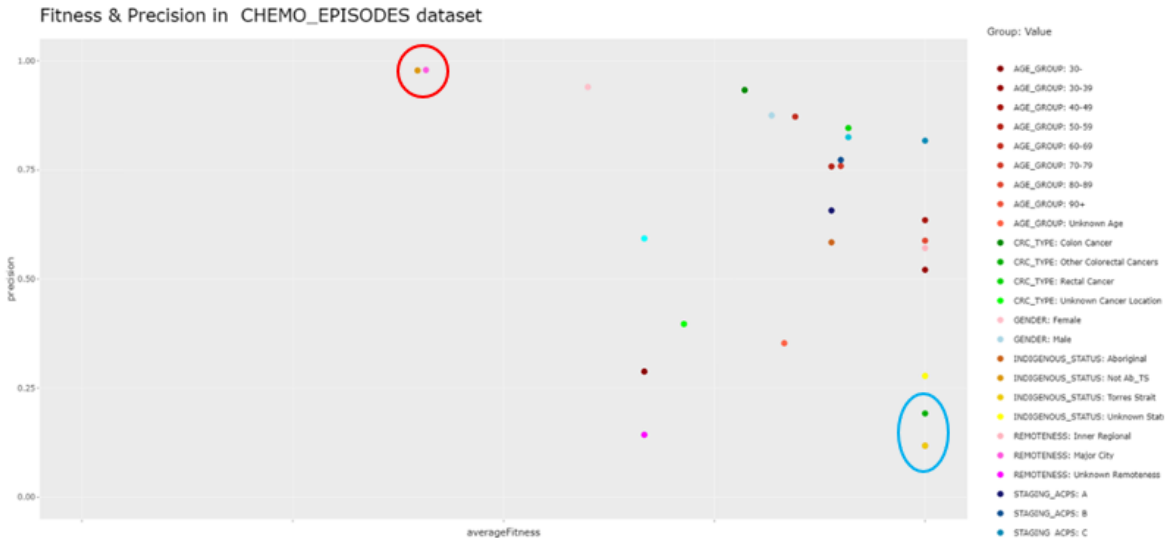



***Figure 30: Fitness and precision of the pathways discovered from dataset 'Chemo episodes'***

We can then zoom in further and focus on the individual metrics, to identify which of the metrics is responsible for lower values. The fitness and precision are then the first group of interest, or the F-score if the quality metric needs to be expressed as a single value. Figure 30 displays the graph of both fitness (how much of the observed behavior of the event log is captured by the model) and precision (how much behavior exists in the model that was not observed in the event log ) for the Chemo Episodes dataset. Groups with a high precision and a low fitness (such as `Major City` and `Not AB/TS`, circled red) have models that do not contain all observed sequences of activities through the pathways, while they do have all individual elements. Conversely, with a low precision and a high fitness (such as 'Other Colorectal Cancers' and 'Torres Strait', circled blue), the model of the pathways contains all elements but also allows for many possible sequences through the pathway than was observed in these subgroups.

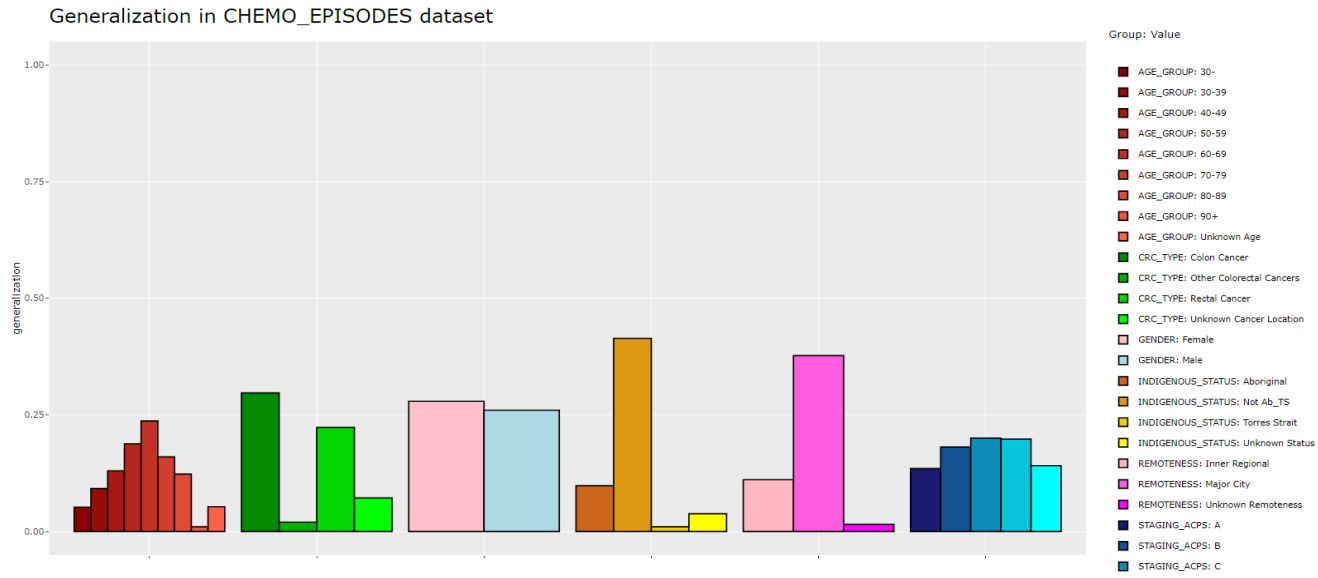



***Figure 31: Generalization of the pathways discovered from dataset 'Chemo episodes'***

The generalization (how well the model explains unobserved system behavior) and simplicity (the complexity of the model) are displayed individually, as the latter is not dependent on the alignments of traces to the model. In the generalization values we observe in Figure 31, we observe that the higher values occur in the subpopulations that are the largest (see **Error! Reference source not found.**a-f), meaning these models could generalize better towards unobserved (or new) activities than their less frequently occurring counterparts. Still, for all groups, the values are relatively low, ranging between 0.01 and 0.5. The models are not suited to display a larger quantity of changes in a pathway, for example, when new regimens are introduced. The models would then need to be updated.

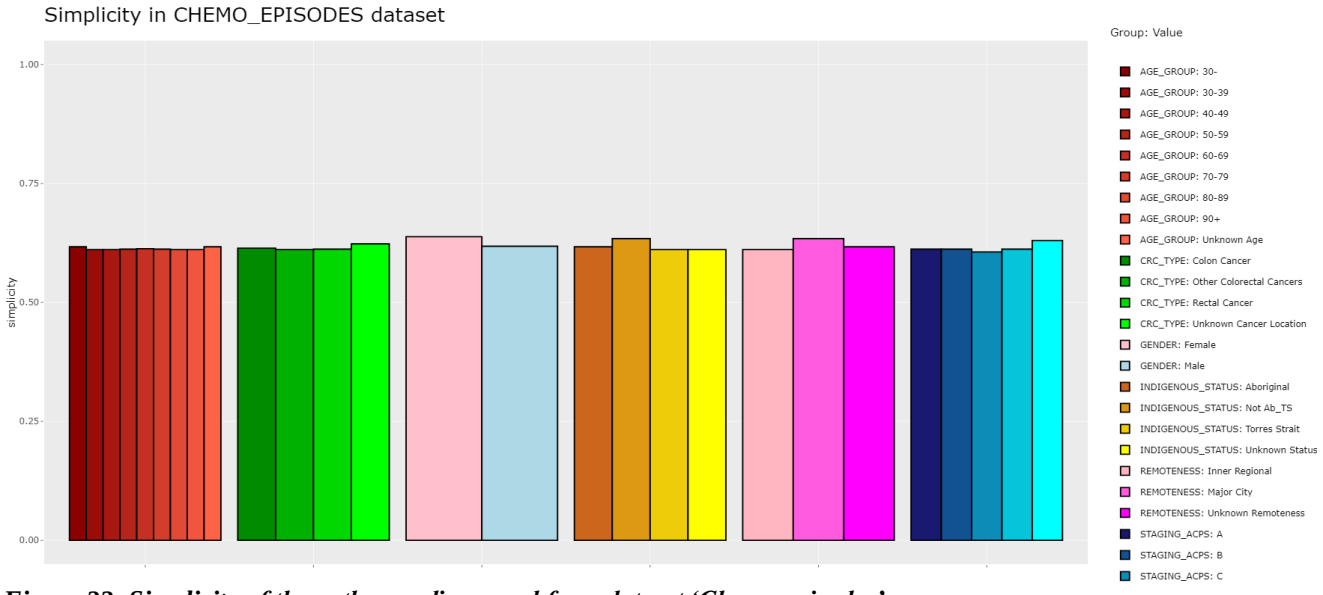



***Figure 32: Simplicity of the pathways discovered from dataset 'Chemo episodes'***

Lastly, we see the simplicity of the models in Figure 32. As expected, these values are nearly identical in all subgroups, as simplicity is not calculated from the aligned traces. That these still differ a little can be explained by the fact that the calculation is made against the number of activities that are in the reference set, so if not all activities in the model are present in the subpopulation's event log, the outcome can be slightly different. With a value of approximately 0.6, this model is not overly complex for explaining the reference data.

#### 5.1.3.1 Conclusion second research question

Returning to the second research question.

> *What are the characteristics of the patients going through the (main) CRC pathways?*

Using the conformance algorithms for alignment-based fitness, precision generalization and for simplicity of the models incorporated in the workflow, we have constructed bar graphs that display the quality of the models, specifically, how well each model represents the patients in a certain subpopulation. From this, we can conclude which subpopulations align the best to this model and answer what characteristics of patients are following the (main) CRC pathways the best.

We have evaluated the quality metrics of the chemo episodes dataset and from the quantitative assessment, we have found that patients going through the main pathways of the phase Chemo Episodes can be characterized as being in the upper middle age groups, mainly the groups 50-59, 60-69 and 70-79 years old. The Age group 30-, 90+ and patients with an unknown age are not close to these values and we can conclude that these groups are not represented well by the resulted model. The pathway is approximately just as good for Colon and rectal cancer, with the populations for 'other locations' and 'unknown locations' less conforming. We see that males on average are better represented by the model, which is mainly due to the higher fitness. The female population has more complex traces, that align less good to the main model. We see that the model best represents the non-aboriginal or Torres Strait islander patients, but that the aboriginal population is relatively close. The Torres Strait islander population does not have a good conformance, even less than the population with an unknown ethnicity. While the number of patients from the Major City is far larger than the rest, the Inner Regional patients are pretty good represented by the model. This suggest there is not a large difference in provided care between those. This is not the case for the unknown remoteness. Lastly, patients in all of the stages, except for the unknown stage are represented relatively good by this model, with the best stages being the ones most frequently occurring.

## *5.2* Cost comparison analysis

To analyse the costs that are incurred in treating the different populations of CRC patients in their entire pathway, including admitted episodes, chemotherapy, diagnostics, GP visits and prescriptions, both Petri nets, created with the custom cost-aggregation algorithm and DFG's created with the ProcessAnimateR package are used. Next to this, density plots of these costs are generated in a hierarchical fashion, first for the entire pathway, then for every phase and then for each of the components in the pathway.

### 5.2.1 Resulting Cost-enhanced pathways

Using the algorithm described in Section 3.2, costs-enhanced pathways are created for all different phases in the dataset. The full pathways have a similar structure as the frequency-annotated graphs in Section 5.1.1 and can be found in the app as well.

First a density plot including a histogram of all the costs in the pathway is created, with an example of the different types or location of cancer visualized in Figure 33. With this visualization, we can see how many patients (on the y-axis) are in different groups of a certain costs, using an logarithmic x-axis. This way we can observe if the costs are skewed differently among the different populations. We see that in this case, we have more colon cancer patients than rectal cancer patients, and that the number of patients with an unknown or other location are negligible. Also, we see that the colon cancer patients have their cost distribution skewed to the right, with a higher average cost.

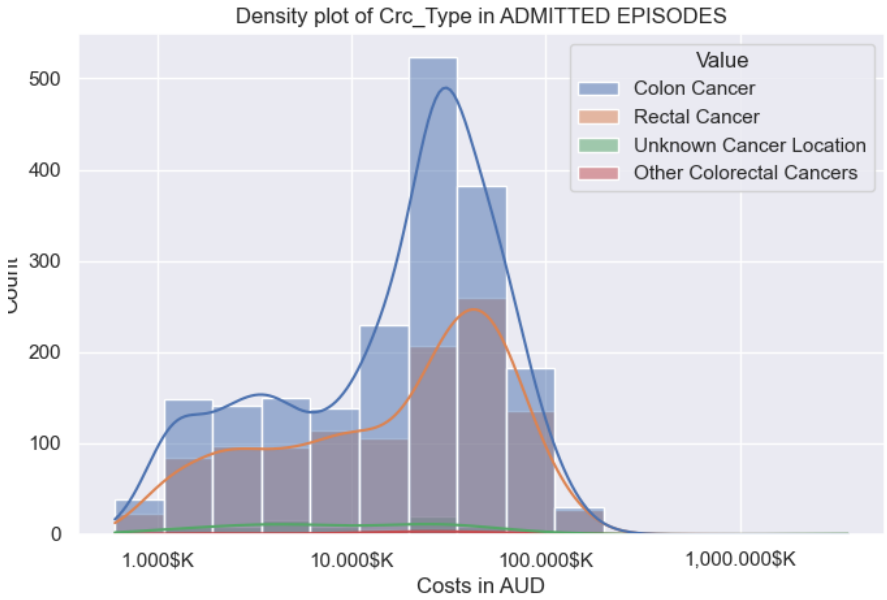



***Figure 33: Density plot of the costs of CRC for each of the registered cancer types in Chemo Episodes***

The following step is to observe where in the pathway these costs are incurred. In the workflow, we have implemented both the version calculating mean costs as well as the sum of all costs. In Figure 34 we see the same part of the pathway of Admitted Episodes as displayed in Figure 19. In this manner, both an average costs of care activities within the model can be visualized. It also gives a more holistic view of the carepathway of the entire population, and where the costs are the highest. In the figure, we see that for example, that while the DRG `PICC` is relatively cheap with an average cost of 977.10 dollar, it still is a significant cost in the entire pathway (1.77M), compared to for example the 24K dollar costing 'Major Small & Large Bowel Procedure' having a total cost of 7M.

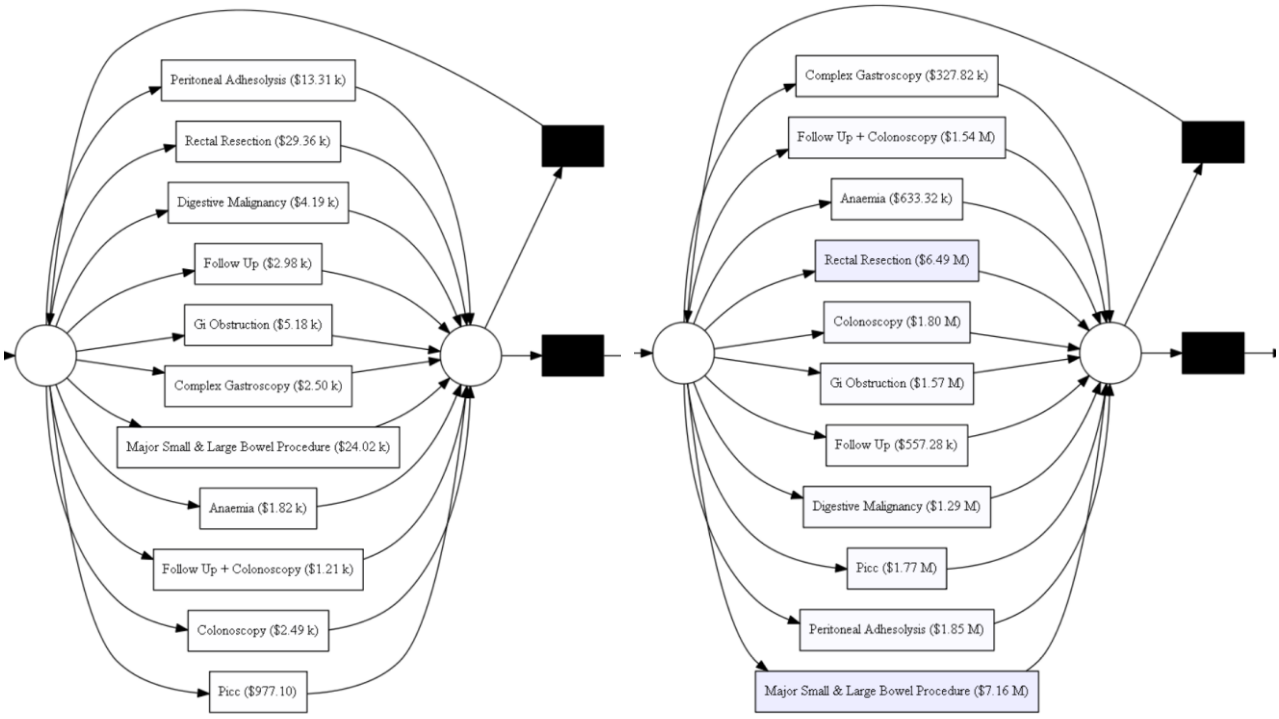



***Figure 34: Component based costs in care pathway of admitted episodes, with mean values (left) and total value (right)***

In the resulting pathways, we can then zoom in into the different care activities and see the distribution of costs within that care-activity, grouped by the same patient-characteristics. In this way we can use the hierarchical structure within our approach, to dive into the different parts of the pathway and analyse individual care activities, all within the same model.

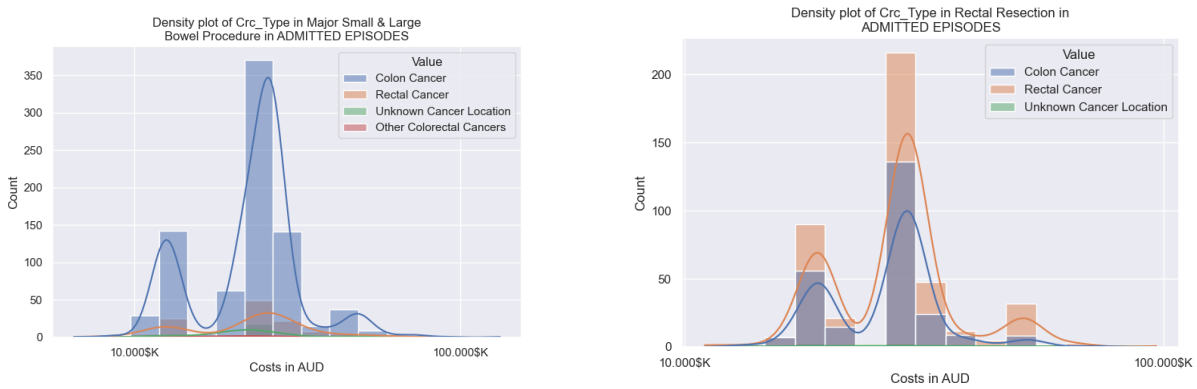



***Figure 35: Individual care-activity cost distributions grouped by patient characteristic Cancer type. Left: Major Small & Large Bowel Procedure, right: Rectal Resection***

### 5.2.2 Process Enhancement with Costs

In the last visualization in the app, the DFG's can be enhanced with this cost information, for each of the datasets where we have these available. In Figure 36 we see an example of this on the Chemo Episodes dataset. Again, in the app the interactive feature is available. In this graph, we can easily identify which components have the higher average costs, by using increasing shades of red for higher values. The individual patients are colored based on the costs that they had incurred up until that point of time, with increasing shading from yellow to red. Note that these values increase and corresponding colors change on the arc before the activity, as the timestamp is linked to the end of the activity.

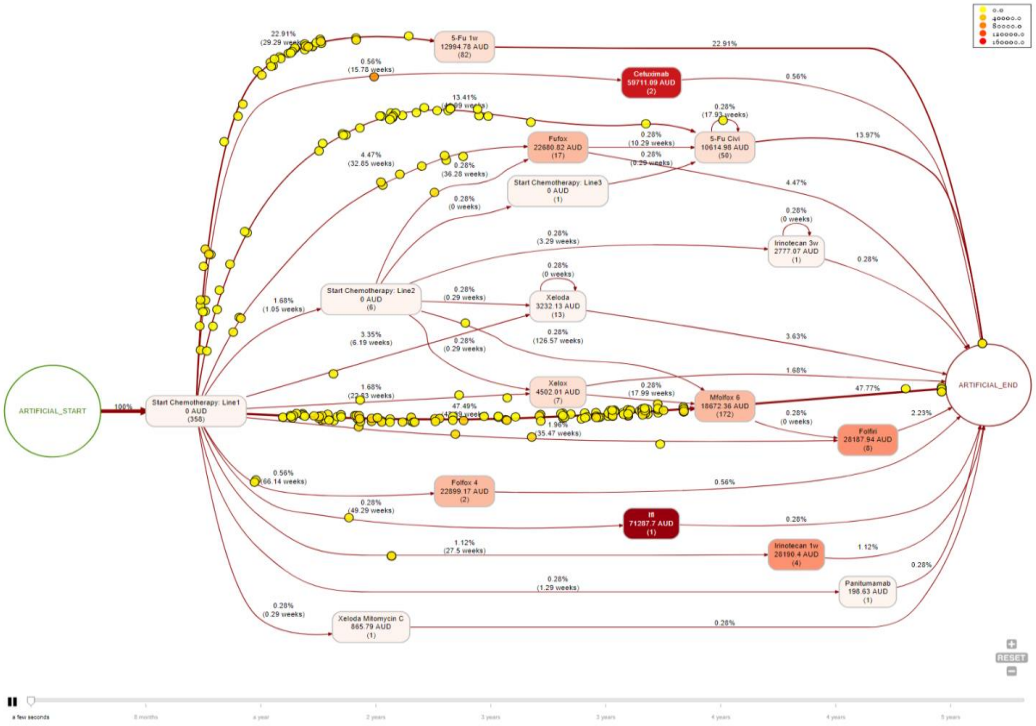



***Figure 36: Direct follow graph of Chemo Episodes, annotated with external data containing the costs of the individual patient up until that point in time.***

Because we built a hierarchical process structure in the chemo dataset bottom-up from individual chemo-medications in each cycle of a regimen, we can also zoom into a level lower than the chemo episode level, to the cycle level. The DFG will be more complicated because of this, but it is still interpretable. We see this displayed in Figure 37, with the notion that the print of the individual regimen names become too small to read. When zooming in, we are able to track the individual patients through their repetition of the cycle, with the additional information on the average cycle duration. We also observe more complexity in this graph, with more episodes connected than was the case within the Episode level model. From this we can conclude that either data pollution has occurred, where some cycles were accidentally registered as the wrong regimen or that within an episode of a regimen, a cycle of another regimen was provided to the patient. This would require further investigation.

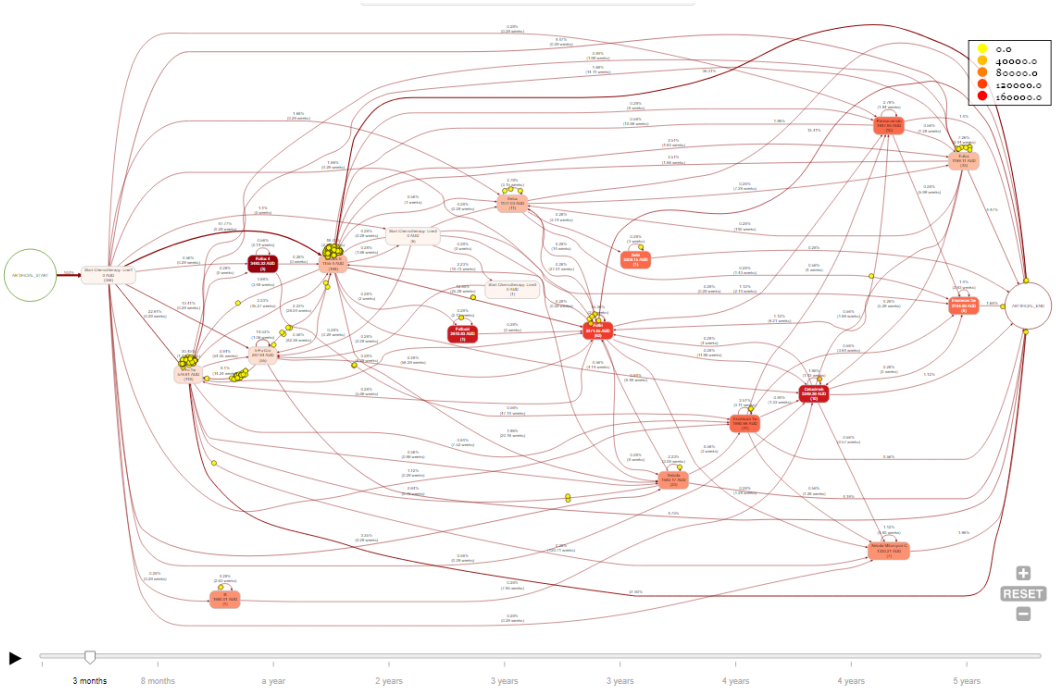



***Figure 37: Direct follow graph of Chemo Cycles, annotated with external data containing the costs of the individual patient up until that point in time.***

### 5.2.2.1 Conclusion third research question

Returning to the third research question:

> *How can the total costs of the (main) CRC care pathways be calculated and evaluated from a cost-perspective?*

Using the developed cost-aggregation algorithm in section 3.2, we are able to calculate mean, and total costs of each of the registered activities in our event logs, by aligning the sequences of activities of each patient over the previously derived process model. These values can be annotated onto the graphs of the resulting Petri nets, giving the observer direct insight into those costs at their respective point in the pathway.

We can then decompose the costs into density plots, comparing the distributions of costs in different subpopulations and do this on each level of our hierarchy, the complete pathway, each phase and each activity in each phase. In this way, differences between the populations can be evaluated iteratively, finding which phase and which activity have more disparities in costs.

Using Direct Follow Graphs, enhanced with the current cumulative costs of each of the patients at every point in their time, we provide insights in the cost-drivers over the life cycle of patients in the health system. When multiple levels of hierarchy in an event log are present, we can evaluate this iteratively to lower levels as well.

# 6 Case Study: Cost disparities between ACPS stages in Colon Cancer

In chapter 5, we showed that we can successfully apply PM to the linked dataset. In this chapter, we aim to assess its use in clinical practice modelling, in this chapter, we evaluate differences between specific groups, to exemplify its purpose and to aid the overall goal in analysing disparities of health care resource utilisation in colorectal cancer. The described methodology and process mining workflow is applicable to all of the registered patient characteristics and associated subpopulations.

We expect from the clinical guidelines for CRC treatment that we will see differences between the care pathways of colon and rectal cancer, and that there will be differences between the lower and higher stages of cancer. We choose to assess the cost distributions and care pathways of the patients in the different stages for Colon Cancer, the larger group of the two cancer sites.

## 6.1 Whole Integrated pathway

First we will evaluate the distribution of costs per patient over the entire integrated pathway of colon cancer patients and find where in the different phases the bulk of the costs are incurred. In Figure 38, we observe that more patients have costs incurred in their admitted episodes, with a distant second being chemo episodes and GP Visits. The total value for n=1965 patients in Admitted episodes is 56,6M (93.34% of total costs). Secondly, the total value for the chemo episodes with n=218 patients is 4M (6.62% of the total costs) The GP visits dataset, with n=95 patients, has a total costs of 14K comprising 0.023% of the total costs. Lastly, Diagnostic tests (n= 30) and prescriptions (n=51) have negligible total costs of 0.01% and 0.006% respectively. The diagnostic tests and the prescriptions have a negligible number of patients. Secondly we see that the largest chunk of the total expenditure is coming from the Admitted episodes, not only by the number of patients that have admissions, but also by their mean costs of approximately 50K AUD per patient. We see that chemo episodes also have a higher mean cost with approximately 30K AUD , while GP visits only cost a few hundred AUD per patient.

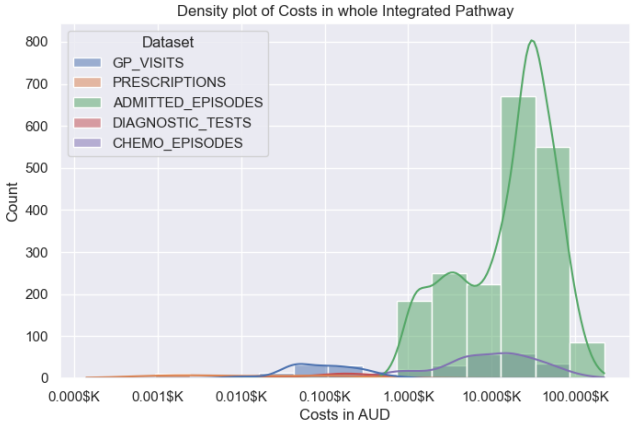



***Figure 38: Cost distributions of Colon cancer in the different phases of care in the whole integrated pathway***

First we look at the average costs of care per patient per stage. We found that the average costs of care for Stage A is 17.808.85 dollars, for Stage B it is 20.988 dollars Stage C it is 27.162 dollars, for Stage D it is 41.643 dollars and for the Unknown Stage it is 10.379 dollars. Following this, we would like to see if these cost distributions are different between the different stages of care delivery. Figure 39 displays a density plot of costs per patient over the entire integrated pathway, grouped by the ACPS stage they are in. We see that Stage B and C patients incurred the highest total costs and that there are relatively small differences between the mean costs in each of the stages, with the mean costs of Stage C and D patients being the largest. We see that the line for patients with an Unknown stage does not have a peak at approximately 50K, suggesting that these patients do not or barely have expensive admissions. The layout of the graph is somewhat misleading, but with changing the x-axis to a scale starting at 1K, we would be able to see differences more clearly.

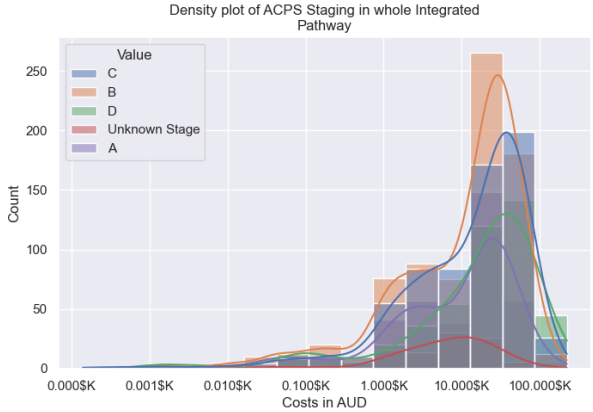



***Figure 39: Cost distributions of Colon cancer in the different stages of cancer over the whole integrated pathway***

A graph of the entire integrated pathway, containing the costs of each of the elements was produced. While the pathway itself is very complex, we can still identify certain regions, see Figure 40. Interestingly, the layout of the frequency annotated version, see Figure 65 in Appendix I, is mirrored, which is probably a result of the automatic layout algorithm. In **Part A**, we see that multiple first line care options possible, mostly containing diagnostic tests such as bloodwork, such as Full blood examination (FBE), Liver Function Tests (LFT) or Urea and Electrolytes (U&E). Additionally, we see in

this part multiple possible prescriptions. Looking at the same pathway, but annotated with the frequency, we see that except for the diagnostic MRI and the admission for an 'Unknown DRG' all the events are unique.
**Part B** contains a similar set of activities as part A, while in **part C** we observe the first chemotherapy regimen starting, as well as an admission for a Minor small & large bowel procedure. **Part D** is for patients who skip Part C, E and F and possibly only have some prescriptions during this period of time. **Part E** solely consists of admissions, of which most are surgery. A small part contains chemotherapy, specifically the MFOLFOX 6 regimen. All activities in this part may be followed by another activity in this part, resulting from the self-loop. **Part F** contains of diagnostic tests, chemotherapy, GP visits and various medication, all of which are mutually exclusive. **Part G** consists of mainly GP visits and prescriptions for medication resulting from that, but also preparation for chemotherapy, which Part G is ended with. In **Part H**, we see only diagnostics and medication. It should be noticed that many of the costs that are calculated with the cost-aggregating algorithm, are zero. This is not because the costs of the activity is zero, but because it was not possible to find an alignment of the traces over the model.

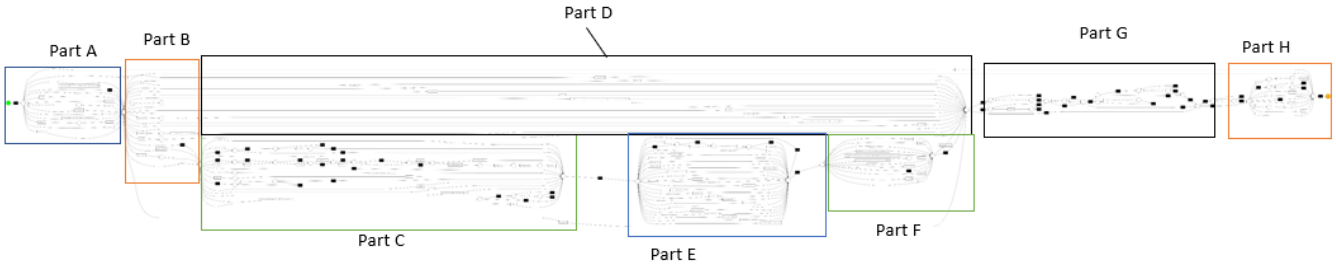



***Figure 40 Entire integrated pathway of Colon Cancer, annotated with costs.***

The process model of the entire integrated pathway is excessively complex and contains many intricate regions. Due to this complexity, the event log is not suited for the described process enhancement, as the creation of DFG's runs into memory-limits. Also, the quality metrics of the model are not very good, see Table 9. These levels are lower in every dimension than we have seen in for Colon Cancer on the chemotherapy set, (see section 5.1.2.1) from which we conclude that the model is overall worse in explaining the dataset than that chemotherapy model was explaining that dataset, mainly lacking precision.

***Table 9: Resulting quality metrics for the entire integrated pathway of Colon Cancer***

| Fitness | Precision | Generalization | Simplicity | Average |
|---|---|---|---|---|
| 0.89 | 0.38 | 0.45 | 0.53 | 0.56 |

From the main model, while we can find regions, there is significant overlap between activities of each of the phases.
To further explore the pathway and find disparities between the stages, we will evaluate the models on the phase-level.

## 6.2 Phase-level Pathways

The pathways with annotated costs are displayed in Figure 41, 43, 45, 48 and 49. We observe that except for the diagnostic test phase, the models have well defined structures. Zooming in on the individual stages, we can display the cost distributions of each of the phases individually, with costs per patient grouped by their stage. In the last of the phases we evaluate, Chemo Episodes, we will dive deep into the differences in costs for each of the ACPS stages.

### 6.2.1 Admitted Episodes

In Figure 41 we see a similar structure as **Error! Reference source not found.** in section 5.1.1, concluding that the colon cancer subpopulation has most influence on the layout of the pathway. We tried to display the enhanced DFG for this event log, but the model is too complex for the memory-limits that the algorithm allows.

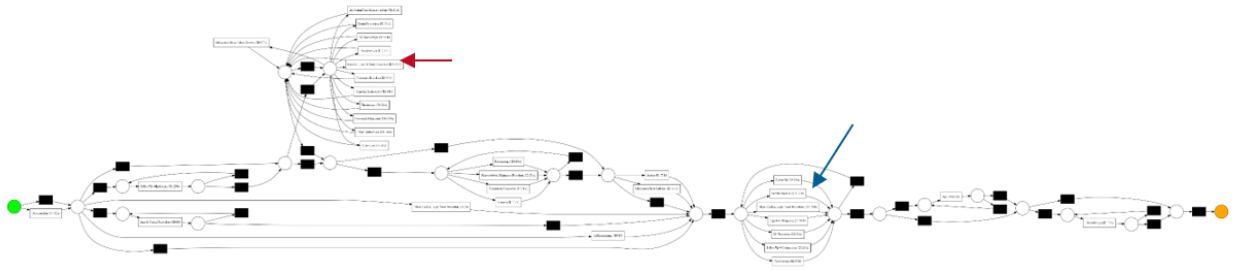


***Figure 41: Total pathways of Colon Cancer with all ACPS substages of admitted episodes***

In Figure 42 we can see that within the admitted episodes colon cancer patients in the lower stage A and B have their distributions more skewed towards the lower end, while C and specifically D are more skewed towards the higher costs. We see from the graph that on average the most costs are incurred in the pancreas, liver and shunt procedure with 29K dollar (red arrow) and surprisingly in the rectal resection 26K (blue arrow). The latter procedure may be performed because the colon cancer had spread to the rectum.

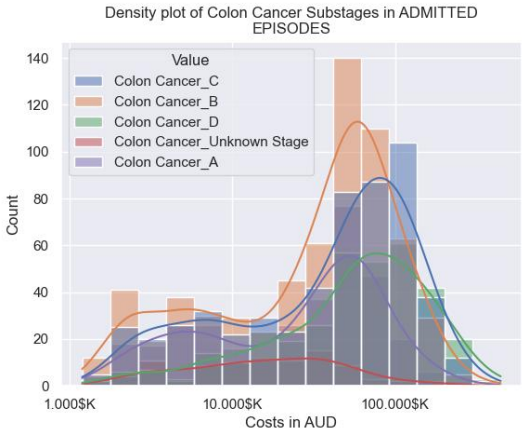



***Figure 42: Cost distributions of Colon cancer in the different substages of cancer in the Admitted Episodes phase***

### 6.2.2 GP Visits

In Figure 43 we observe four instead of three sections, compared to the structure of Figure 23 in section 5.1.1. In the end of the pathway, the multidisciplinary team meeting has become a separate part. However, as the costs associated to it that are zero, we note that there were no traces aligned correctly to the model. Looking at the distribution per stage Figure 42, we observe that the costs for all stages are quite similar, while it should be noted that the total number of patients having GP visits is low and that because of this, distributions might not be distinguishable. In the Petri net, we see that the highest cost are incurred by the 'Other category ' (red arrow), although we know from the DFG that the health assessments were actually more expensive, costing 196.25 dollar each on average.

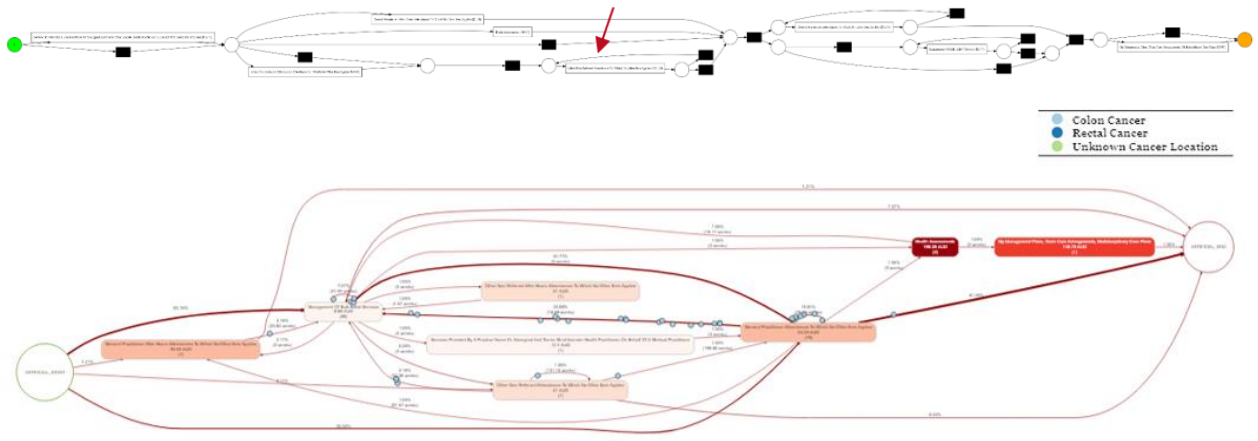



***Figure 43a&b Total pathways of Colon Cancer with all ACPS substages of GP Visits***

In the cost distribution of Figure 44 we see that the costs of the GP visits are relatively equally distributed in all of the phases, but we make notion of the fact that the number of patients per bucket is so small, that this may be misleading. We would expect differences in the number of GP visits between the stages, and while we see more visits for lower stages, especially stage B, we have recorded not that many overall, making the distinction between them hard.

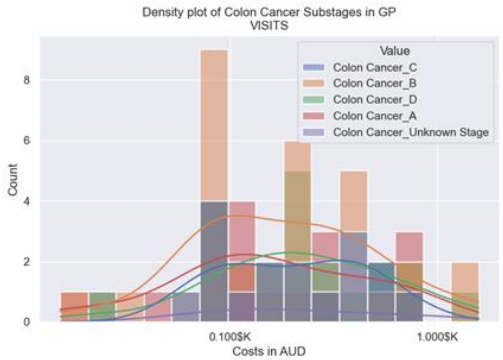



***Figure 44: Cost distributions of Colon cancer in the different substages of cancer in the GP Visits phase***

### 6.2.3 Prescriptions & Diagnostic tests

In the Figure 47 we see the Petri net of the prescriptions. The DFG was not able to be generated due to the size of the net. In contrary to the model, we saw in Figure 24 in section 5.1.1, we see more sequences and fewer concurrent prescriptions. We can see in Figure 46 that almost all the group sizes have only one observation. The sequences we see in Figure 47 may be the result of the fact that the algorithm processed many unique traces and therefor is unable to generalize well.

***Figure 47: Total pathways of Colon Cancer with all ACPS substages of Prescriptions***

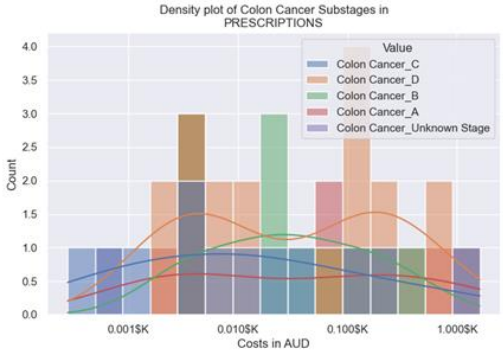



***Figure 46: Cost distributions of Colon cancer in the different substages of cancer in the Prescriptions phase.***

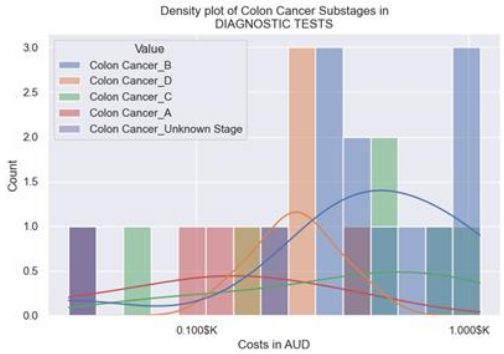



***Figure 45: Cost distributions of Colon cancer in the different substages of cancer in the Diagnostic Tests phase***

In the pathway for diagnostic test Figure 48, we do see a similar model as in Figure 24 in Section 5.1.1. Figure 45 shows us a cost distribution which similarly consists of mostly unique observations. This results in distributions that are nearly uniform and do not represent the underlying distribution very well, because it is very susceptible to outliers.

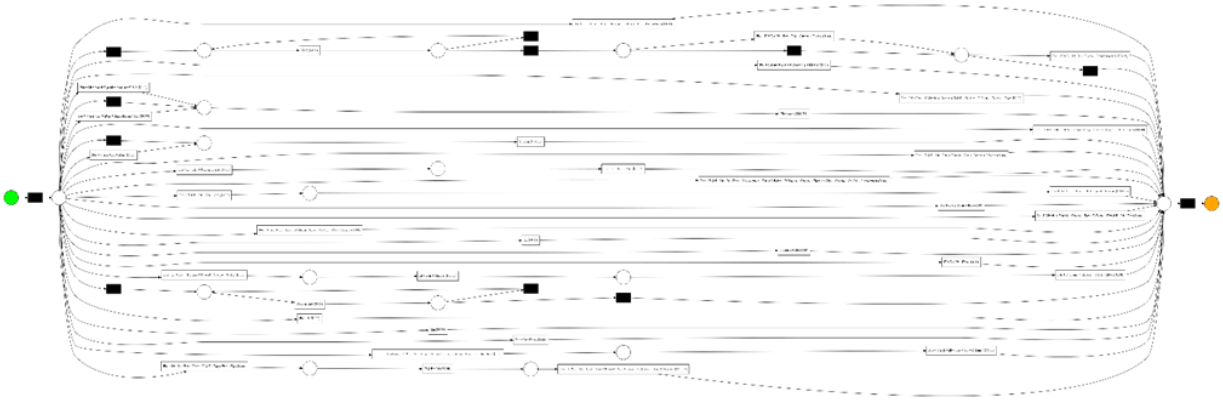


***Figure 48: Total pathways of Colon Cancer with all ACPS substages of Diagnostic Tests***

### 6.2.4 Chemo Episodes

The petri net of the model for chemo Episodes and the associated DFG, see Figure 49, are much more distinctive than in the two previous phases. In the Petri net we can clearly see the first line and second line regimens. Interestingly we see that the second line treatments, MFOLFOX 6 and FOLFIRI are not concurrent, but rather in a sequence, while there was no patient that had those two regimens after each other. We know this, because there are only two lines of chemo treatment started and that it was not possible to skip the middle section. In both the Petri net and the DFG, we observe that the start activity of the second line is earlier than the actual end of the first line. This is unexpected, but can be explained by the fact that the timestamps of the beginning of the second line coincide with the end of the first line.

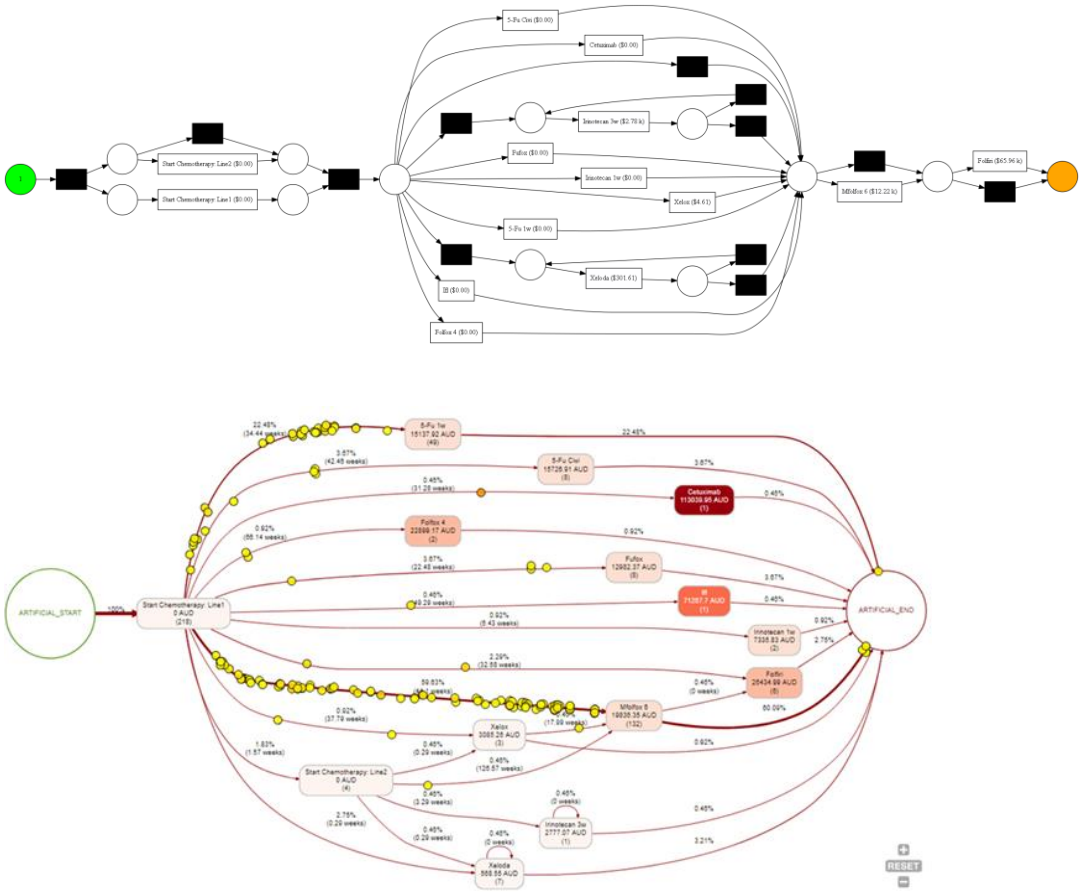


***Figure 49: Total pathways of Colon Cancer with all ACPS substages of chemotherapy***

First we look at the average costs of care per patient per stage for chemo therapy. We found that the average costs of care for Stage A is 25.007 dollars, for Stage B it is 21.025 dollars, for Stage C it is 11.227 dollars, for Stage D it is 23.295 dollars and for the Unknown Stage it is 9.887 dollars. The relatively low average costs of stage C patiens compared to B & D is unexpected. When we look at the cost distributions of the different stages in chemotherapy (see Figure 50), we see a clear difference between the mean costs of stage C and stage D, while Stage A and B are somewhere in between. As expected from the clinical guidelines, Stage A patients hardly have chemotherapy. We would expect the higher stages to have more and more expensive chemo-therapy, and we observe that this is the case for stage D patients. However, we see that the density of stage B patients is skewed more towards higher costs than stage C, which might be explained by the longer duration of patients receiving chemo therapy or towards the fact that recurrent and metestatized cancers are still categorized based on their primary staging.

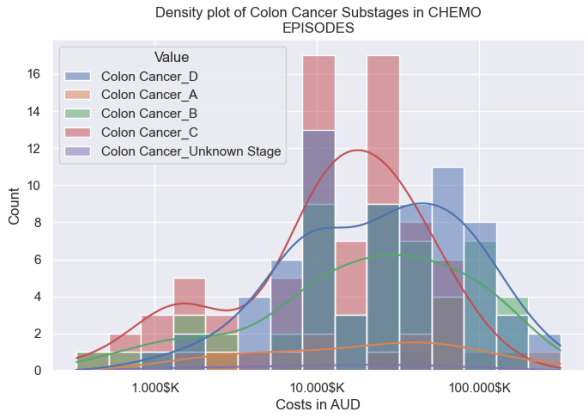



***Figure 50: Cost distributions of Colon cancer in the different substages of cancer in the Chemo Episodes phase***

To further explore what kind of chemotherapy regimens are the main cost drivers for each of the stages, we will evaluate the pathways on the individual characteristic (ACPS stage) level.

## 6.3 ACPS stage-level Pathways

In the last part of the evaluation, we will dive deep into the chemotherapy phase and compare the DFG's of each of the ACPS stages, annotated with their costs. In Figure 51 we see the graph of Stage A, having a total of 14 patient with a total costs of 350K, where we see only four possible regimens. On average, the costs of MFOLFOX 6 is the most expensive, having an average cost of approximately 35K dollar. Cheapest is the regimen provided to one patient, with FUFOX. Only one of these patients received two lines of chemotherapy, two times with a relatively cheap regimen of Irinotecan 3W.

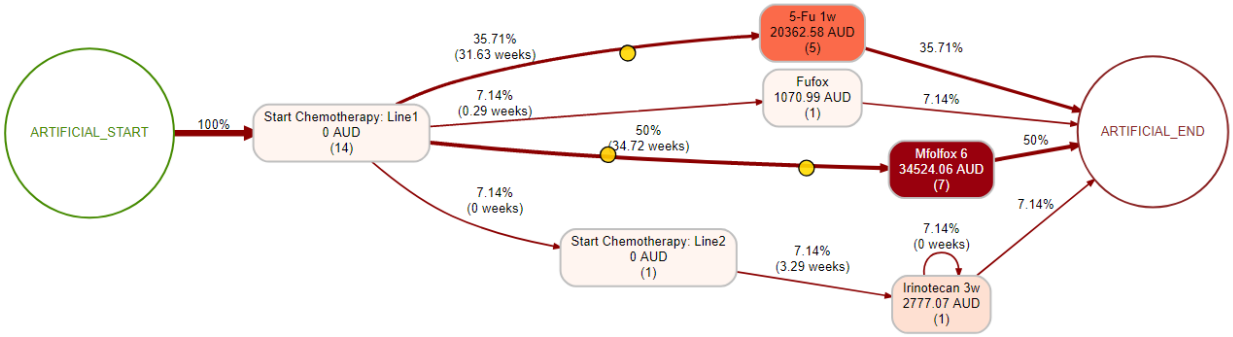



***Figure 51: Direct follow graph patients with Stage A (N=14) Colon Cancer in Chemotherapy***

In the graph of stage B, Figure 52, we have 53 patients with a total costs of 1.11M dollar, and we see a lot more regimen options. Eye-catching are the red nodes, with exceptional high cost for the two patients receiving FUFOX (61K dollar) and IFL (71K dollar). Compared to the patient in stage A, the FUFOX regimen is much more expensive, which explains the skew in the distribution of Figure 50 toward higher costs. The more frequently provided regimens, MFOLFOX 6 and 5FU-1W are more representative of the expected costs for stage B Colon cancer patients.

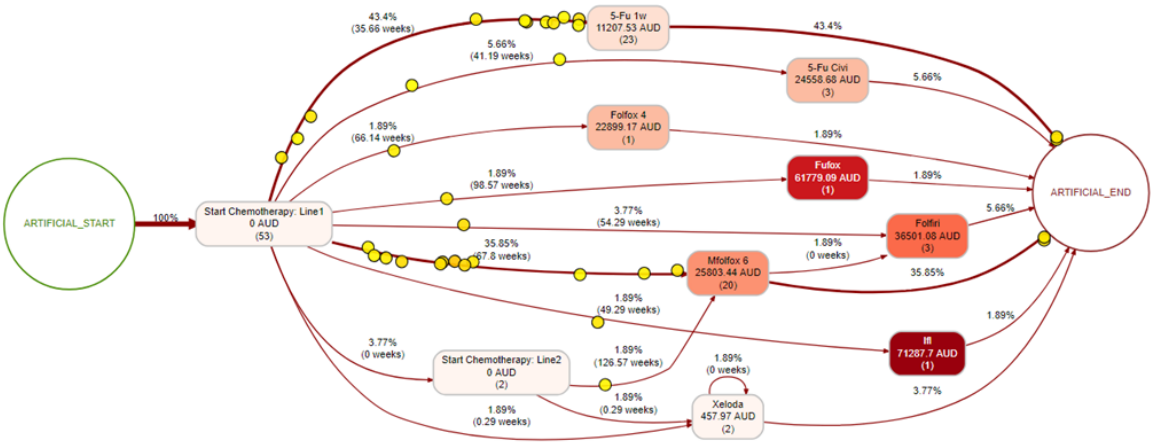



***Figure 52: Direct follow graph patients with Stage B (N=53) Colon Cancer in Chemotherapy***

In Figure 53, we have 77 Stage C patients with a total costs of 865K, that all only have first line chemotherapy. Again, there are outliers, with a single patient receiving FOLFOX 4, but mostly, the costs are between 10 and 17K per regimen, with the exception of cheaper regimens for Xeloda and Irinotecan 1W.
The large number of patients receiving the relatively cheap MFOLFOX 6 regimen, make up the most of the costs, explaining the skew from the distribution in Figure 50.

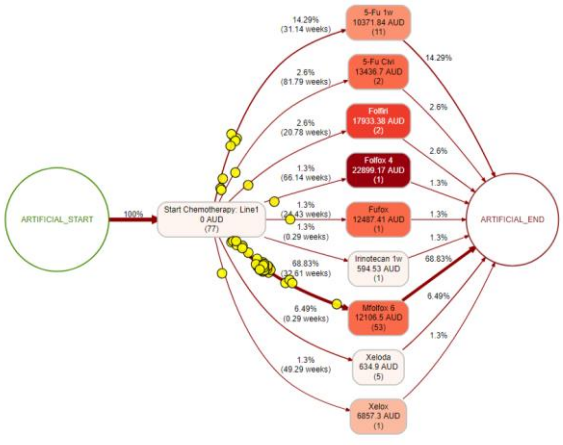



***Figure 53: Direct follow graph patients with Stage C (N=77) Colon Cancer in Chemotherapy***

In the highest stage D in Figure 55, we see another 71 patients with a total cost of 1.65M dollars, having one outlier receiving Cetuximab against a cost of more than 110K. Again, most patients receive MFOLFOX 6 and we see comparable costs for 5-FU 1W (9 patients) and FOLFIRI. We see one patient having MFOLFOX in the second line, after starting on XELOX, but this was discontinued after 3 days. Again, the costs for the three patients receiving FUFOX was low compared to FUFOX receiving patients in stage B and C.

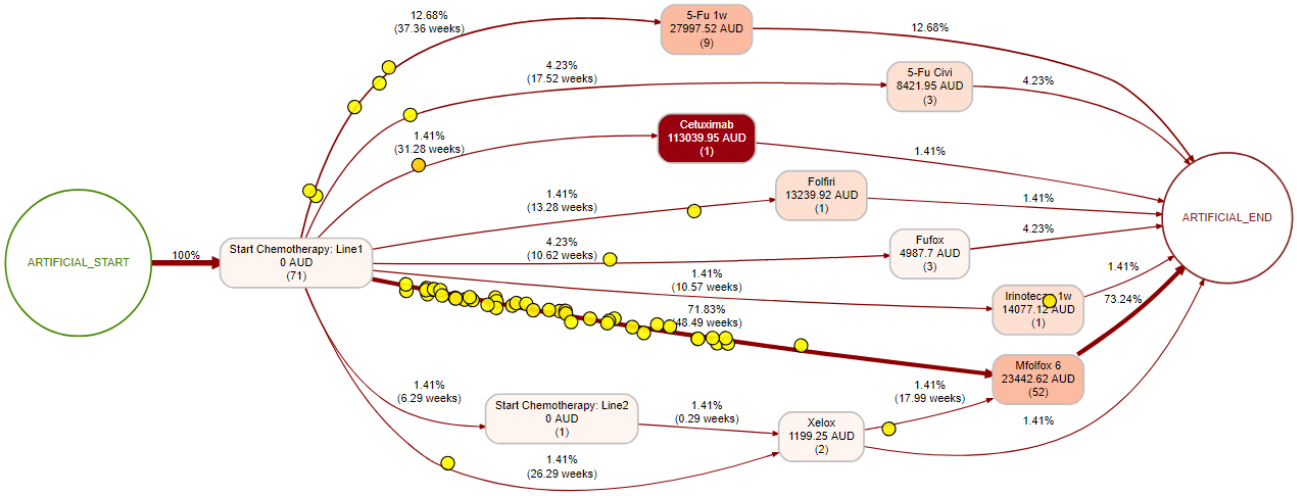



***Figure 55: Direct follow graph patients with Stage D (N=71) Colon Cancer in Chemotherapy***

Finally, we have 3 patients where no stage was registered, with a total costs of 29.6K. In Figure 54 we see that they all had one line of chemotherapy, either an 5FU 1w for 16K dollar or the FUFOX regimen, for 7K dollar. The costs of these patient are comparable to Stage A patients. In the total costs incurred in Chemotherapy, these numbers are negligible as well.

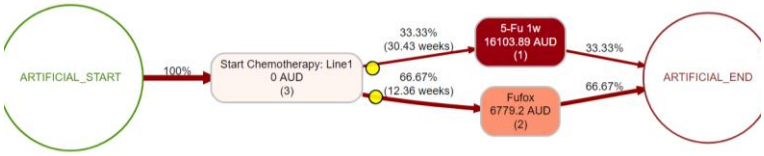



***Figure 54: Direct follow graph patients with an Unknown stage (N=3) Colon Cancer in Chemotherapy***

As we have seen, the MFOLFOX 6 stage is most prevalent in each of the pathways and seems to have different costs for the different phases.
Figure 56, displaying the cost distribution of all the stages in MFOLFOX 6 confirms this. We see that stage D patients have much higher costs for MFOLFOX 6 compared to stage C patients, while stage B patients, even though there are far fewer, have a more skewed distribution towards higher costs. Stage A patients have a widely spread cost distribution, but a minor influence on the total costs incurred in MFOLFOX 6, as there are a negligible number of them.

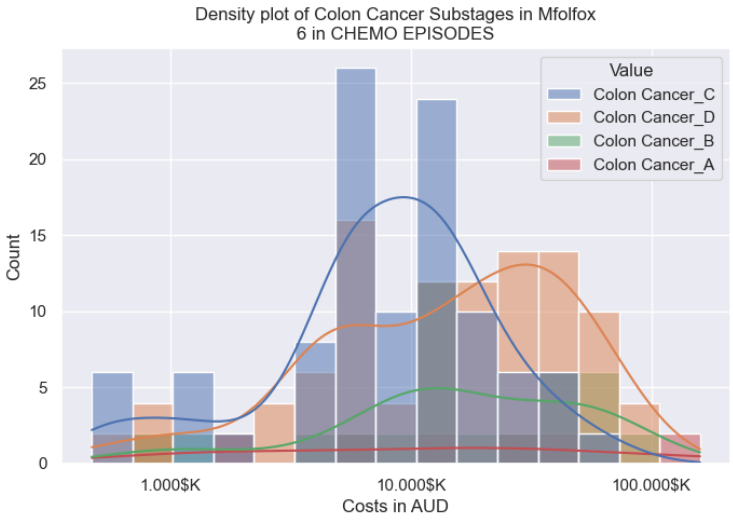



***Figure 56: Cost distributions of Colon cancer in the different stages of cancer, that received regimen MFOLFOX 6 in the Chemo Episodes phase***

## 6.4 Conclusion Case study

Over the entire pathway, we observe that there are small differences between the distributions of costs for each of the stages of Colon Cancer. We have seen that the admitted episodes contain the largest part of the costs, and that the lower stages stage A and B have their distributions more skewed towards the lower end, while C and specifically D are more skewed towards the higher costs. The Chemotherapy phase has a clearer distinction in costs distribution between the stages, while within the other phases, the distributions are not very different. A note in this regard is that the absolute number of patients in a phase or stage can heavily skew the distribution and also misrepresent the calculated mean. For Chemotherapy, we can conclude that most of the costs incurred in this phase come from MFOLFOX 6 and that the large differences per stage in this phase are mainly due to the large differences between stage D patients receiving MFOLFOX 6.

# 7 DISCUSSION

This chapter discusses the results of the previous chapter, answers the formulated research question from section 2.6 and discusses implications for clinical care, limitations in this research, future directions of Process Mining research in health services and recommendations.

## 7.1 Conclusion on the main research question

In the previous chapter we have seen the results of applying the designed workflow to the entire selected cohort of CRC patients. Moreover, we have conducted a case study evaluating the cost differences between patients in different stages of CRC. Coming back to our main research question, we combine the insights from our conclusions:

*"How can Process Mining be applied to derive care pathways and analyse the costs of care provided to CRC patients in these care pathways?"*

We have described a workflow to create event logs suitable for Process Mining, using linked data from multiple Australian Colorectal Cancer registries. The workflow derives process models for the carepathways of the main phases in colorectal cancer care. This workflow contained all three process mining dimensions: Discovery, Conformance and Enhancement. In the discovery parts, the inductive miner algorithm was used to derive the process models, that could be visually inspected and interpreted. Additionally, from the event logs Direct Follow Graphs were constructed, which could function as an enhancement of the visualizations, displaying characteristics on a patient-level. As well, these graphs can be extended with external event logs, containing information of patients that change over time, so they, for example, provide context on the patient's status or the patients treatment line or regimen.

The equivalence of pathways can partially be evaluated by comparing the resulting alignments of the models visually and this can be extended by quantitative analysis using the conformance metrics, resulting from the conformance algorithms in the workflow. A combination of both is needed to adequately evaluate the pathways that result from the process mining workflow. The conformance algorithms for alignment-based fitness, precision generalization and for simplicity of the models incorporated in the workflow. From the outcomes of these algorithms, bar graphs are made that display the quality of the models, specifically, how well each model represents the patients in a certain subpopulation. From this, we can conclude which subpopulations align the best to this model.

Using the developed cost-aggregation algorithm in section 3.2, we are able to calculate mean and total costs of each of the registered activities in our event logs, by aligning the sequences of activities of each patient over the previously derived process model. These values can be annotated onto the graphs of the resulting Petri nets, giving the observer direct insight into those costs at their respective point in the pathway. We can then decompose the costs into density plots, comparing the distributions of costs in different subpopulations and do this on each level of our hierarchy, the complete pathway, each phase and each activity in each phase. In this way, differences between the populations can be evaluated iteratively, finding which phase and which activity have more disparities in costs. Using Direct Follow Graphs, enhanced with the current cumulative costs of each of the patients at every point in their time, we provide insights in the cost-drivers over the life cycle of patients in the health system. When multiple levels of hierarchy in an event log are present, we can iteratively evaluate this on lower levels of granularity as well.

Our case study showed that for Colon Cancer Patients over the entire pathway, that there are small differences between the distributions of costs for each of the stages. We found that the average costs of care for Stage A is 17.808.85 dollars, for Stage B it is 20.988 dollars Stage C it is 27.162 dollars, for Stage D it is 41.643 dollars and for the Unknown Stage it is 10.379 dollars. The admitted episodes contain the largest part of the costs, and that the lower stages stage A and B have their distributions more skewed towards the lower end, while C and specifically D are more skewed towards the higher costs. The Chemotherapy phase has a clear distinction in costs distribution between the stages, while within the other phases, the distributions are not so different. Unexpectedly, stage C patients had less expensive chemotherapy then stage B & D patients. We found that the average costs of care for Stage A is 25.007 dollars, for Stage B it is 21.025 dollars Stage C it is 11.227 dollars, for Stage D it is 23.295 dollars and for the Unknown Stage it is 9.887 dollars. A note in this regard is that the absolute number of patients in a phase or stage can heavily skew the distribution and also misrepresent the calculated mean. For Chemotherapy, we can conclude that most of the costs incurred in this phase come from MFOLFOX 6 and that the large differences per stage in this phase are mainly due to the large differences between stage D & B versus stage C patients receiving MFOLFOX 6.

## 7.2 Limitations

This study into the use of Process Mining has several limitations. First, the outcomes of the models have not been clinically validated, so their usefulness for clinicians is not yet known. It is expected that validation with clinical input will be needed to establish the interpretability and added value of the obtained models. An iterative approach to improve the data cleaning and wrangling, deriving and quality checking of models and then interpretation and valuation of the models by clinicians would be preferable.

Another limitation to this research is that the used linked datasets contain patient data captured in electronic health records of a subset of hospitals in the Melbourne metropolitan area, and for the first line treatment only from General practitioners collaborating with Medicine Insight. These potentially does not cover the entirety of care provided in all health facilities to all patients, leading to possible bias in the pathways. Also, this data reflects clinical practice for patients ultimately diagnosed with colorectal cancer, and therefore can only be used to assess changes in this specific patient group, while the population with a suspicion for colorectal cancer would be larger.

Bias can also arise from the imbalance in the number of patients in each of the resulting datasets. The entire population contains 4246 patients, and most of these patients (3233) were linked in the VAED dataset. However, the number of linked patients in the primary care dataset was only 1105, and after filtering their encounter reasons based on a list of symptoms for CRC, only 187 unique patients remained. While the distribution of patient characteristics is comparable to the main cohort, there is a fair chance of selection bias.

We observed that the resulting names for the elements in the primary care dataset were mostly unique, as they were derived from free-text fields in the dataset and were not clustered into less granular groups. The uniqueness of all care activities resulted in pathways that could not display concurrency and other sequencing relations, leading to less informative models. This limitation can be overcome by either a new classification system in the raw data collection as well as a form of classification stemming from the field of natural language processing (NLP).

The resulting models depend on the specific discovery algorithm used, as well as on parameter settings within these algorithms. In Section 3.1 we observed the differences between the inductive miner algorithm and the heuristics miner algorithm, which vastly differs in interpretability. The first allows for loops and it guarantees sound process models, and it is expected that the resulting model structure will be different with other discovery algorithms, which might result in different outcomes for the conformance metrics or problematic conformance checking.

We also observed some unexpected behavior in the DFG's, with regards to sequencing. As the timestamp of finishing one line of chemotherapy coincides with the timestamp of starting another line, the algorithm placed the start of the second line mark before the end of the previous line. This limitation can be diminished by explicity coding the sequence or by manipulating the timestamp slightly, for example adding a few seconds to the start moment, in order to erase this problem.

A last limitation is with regards to the information provide by the cost-aggregation algorithm itself. Currently, this algorithm does not incorporate an output giving an measure of what percentage of the costs is unattributed to one of the activities in the pathway. Without this, the 'unexplained' costs within a pathway is not known, while this would be valuable information for the researcher, to conclude how good the quality of the cost-aggregation by alignments itself is. An extension to this algorithm is therefore recommended.

## 7.3 Impact and relevance for clinical care

Evaluations of clinical practice play an increasingly important role in managing and improving quality of care. The data-driven approach of PM in modelling clinical practice may increase the accuracy with which clinical practices are known and can be represented. In Health Services research, clinical guidelines and the optimal care pathway models obtained from consensus-based meetings are used to describe care as it should be provided. In contrast, the models obtained with Process Mining methodology are a description of care as it is actually in practice provided. The optimal carepathway models do not and cannot capture the evident complexity of care in practice in a reasonable amount of time, let alone compare the differences between care provided between certain subpopulations in a data-driven and quantative manner. Decision-makers and health services researchers that use these OCP models to guide their choices in designing care delivery would benefit from a more accurate depiction of current care. This would improve subsequent (simulation) modelling for introducing new technologies or health services design, as well as a more complete retrospective quality assessment of the entire integrated pathway.

A large advantage of these automatically derived process models is that they do incorporate the deviations doctors make and give a more complete representation of the care pathway. Being able to automatically derive process models for complex care pathways can greatly reduce the amount of work needed to find disparities between groups of patients. Process mining applied to health services research context allows easy and frequent checks of the pathways, which is beneficial with the analysis of adoption of newly introduced guidelines or innovations. It can also aid in identification of regional differences in adoption of changes in clinical practice and it allows for identification of subgroups of patients or parts of clinical practice where care delivery is different. Partington et.al. presented preliminary findings from a case study of comparative process mining that utilizes routinely collected data to describe differences in the process of care as delivered at four Australian hospitals [91].

In the context of health economic modelling, Process Mining can aid in the early stage to derive a model structure on which a simulation model can be build. Health economic evaluation can be based on cohort-based state transition models as well as discrete event simulation (DES) that is increasingly used in health economic evaluations, to implement more complex model structures [92]. The mathematical properties from Petri nets allow for simulation purposes and van der Aalst et. al. proposed an outlook on the combination of retrospective Process Mining, from which prospective simulation models can be derived [93].

The rapid development of Process Mining in the healthcare field may result into adoption of the techniques as a standard tool for evaluating care provision, even on a real-time basis, by healthcare professionals and health services researchers. Ideally, the workflow for process mining will be applied to streaming healthcare data from an electronic health record system where hospitals and primary care providers register the necessary information. The structure for analysis of the outcomes may be based on the value-based paradigm around the integrated pathway, providing real-time insights of patient's routes through the healthcare system, including deviations from the expected pathways, statistics on the performance of the hospitals with regards to duration and costs accumulated in the pathway. Clinicians may then focus on the patient outcomes and quality of care provided, while healthcare managers may focus more on the costs and the performance of the system.

## 7.4 Challenges for PM in health services research

As seen in the models of Figure 24, displaying the prescriptions and diagnostics, models derived from event logs containing mostly unique events do not provide the same level of insight as models derived from logs with frequently reoccurring activities. With many unique events, most sequencing concepts that Process discovery can derive, cannot be found. Concurrency, successive and choice relations are not visible in many unique pathways, unless the order of the activities is exactly the same with each case. A solution to this type of problem is to introduce more hierarchical clustering of activities with comprehensive group names and then iteratively mine process models on each level of the hierarchy. For example, within the prescriptions, all medication related to a certain symptom could be clustered, for example, 'nausea-related' medication. Still, it needs to be considered that within these groups where only a choice relation exists, the additional value of PM for control-flow would diminish.

An additional challenge in this regard is that hierarchical clustering of events currently is a supervised (i.e., manual) task and requires a certain minimum level of insight in the medical domain. By using administrative events from electronic patient records, changes in events can be related to changes in logistics in addition to actual changes in clinical care. Additionally, when changes occur to how events are registered, terms are updated, or the rule-based clustering strategies change, the discovered process models will also be different. However, general noise-reducing algorithms are available that allow partial automation of the selection and clustering of events, mitigating the last limitation [61]. Furthermore, developing a standard (disease-specific) classification scheme and clustering strategy for events may overcome this challenge in the future.

To obtain more insightful outcomes of PM, the quality of registered events is important. Events need to be properly transcribed as an identifiable activity, as changes in the model can only be identified from the information that is captured in the data. A more accurate representation of clinical practice can be achieved with a more extensive event log, which makes use of diagnostic results, e.g., lab test results, CT-scans and clinical characteristics of the patient. This can be incorporated into data-aware petri-nets (DPN), a form of petri-net that uses additional information at decision points in the process model [29]. The usage of DPN in itself would not change the methodology and workflow, but the resulting graphs will add value to clinical practice, with the notion that this also add to the model's complexity.

Another challenge in PM projects regarding complexity is that of the algorithms involved. The computation of the alignments has $\boldsymbol{O}(b^d)$ complexity, where $b$ is related to the number of unique traces in the set and $d$ is related to the number of unique activities in the traces. As observed in the experimental runs, the computation time of the larger dataset Admitted Episodes, containing approximately 3200 patients with a maximum of approximately 50 unique activities, a single run of the alignments took approximately 1100 seconds, resulting in a total of 26 hours runtime for the entire pipeline. Compared to the half an hour run for the pipeline of the Chemo Episode dataset, containing 461 patients with a maximum of approximately 15 activities, this is excessive. Even though distributed computing through cloud platforms vastly improves the computational power, the event log sizes can only increase a few orders of magnitudes before this also becomes infeasible. When commencing a PM project and constructing the event log from EHR data, the number of unique traces and unique activities should be estimated and brought back by hierarchical clustering.

Lastly, as Process Mining is a young field and growing rapidly, the software used can become outdated fast. Within this project, we have encountered updates to the PM4PY package in which some functionality was deprecated over the duration of the study. We also see that the quality of the data visualizations differs vastly between packages. The fast-paced improvement of software makes it hard to standardize code and create durable data pipelines. When the developments cycle will reach maturity, this challenge may be overcome, but until that moment, researchers should expect to maintain their code extensively.

## 7.5 Future directions

Process Mining has the potential to diminish the workload needed to model and analyse complex care. As demonstrated, the work can be applied to linked data from multiple sources and automatically derive models that can be used to analyse care provided to patients in multiple phases of care provision. In conducting this study, we identified challenges that can pose starting points for additional research.

The first direction to be explored further is into the pathway of primary care. Previous research of process mining in oncology is usually based on hospital care. In this research, the quality of the data of the NPS dataset was somewhat lacking and the total amount of patients linked to the registry was low. This led to process models with almost unique pathways for each patient. When a dataset with a larger number of patients is obtained and more clear naming conventions are chosen, the pathway of primary care will provide more insight into the lead-up to a diagnosis. Data quality in this regard is important, but also a better way to include registered activities, medication, visits and diagnostics than basing it on text recognition.

Secondly, in the current study, we have enhanced the process models with cost information and following the value-based healthcare paradigm, the health services research field could benefit from analysis on the outcomes of patient's as well. Insights in the level of wellbeing of patients throughout their care pathway, as patient-reported outcomes are quantified and could be used as additional enhancement of these pathways, in the same fashion that the costs have been added in this project. When researching this data-aware petri-nets could be of value.

Another concept for further investigation is on the variation of clinical practice over time. Care delivery is not static and can continuously change due to new insights and the implementation of new innovations or new design of the system of health care delivery. There has been some (yet to be published) research in this field. [94] Analysis in change in care delivery over time can give insight in whether certain treatments are stable over time, such as routine procedures that are not patient specific or that they change frequently due to new technology or designs of systems. Also, time-variation analysis can identify subgroups where care processes are erratic, indicating a potential need for further guidance (guidelines) or further implementation strategies.

The last future direction that can be further explored is the comparison of guidelines to actual clinical practice, including the use of linked data. Guideline compliance checking is a subfield where multiple studies have case studies for, but they tend to use only single instances of electronic health records and mostly evaluate pathways within a single hospital or hospital group. It would be interesting to evaluate the guidelines between suspicion and treatment, focussing on the now not yet known pathways of assessments, diagnostic tests, and referrals.

## 7.6 Recommendations regarding Process Mining in HSR

Subsequent research using Process Mining can benefit from the insights obtained in this study. Regarding *the methodology*, we state the following recommendations:

- Start with a validation of the resulting pathways by clinicians, looking for unexpected situations or difficult to interpret models. The additional value of the observed models for clinical care is not yet validated, as well as its applicability to other disease types, or health services research in general.
- Secondly, extend the cost-aggregation algorithm with internal diagnostics on the unexplained fraction of the costs. This will give the researcher better insight in how well the algorithm performed on the evaluated pathway.
- When commencing a Process Mining project, the researcher should start with setting the basis of which kind of activities the mining should be performed and estimating the number of unique activities they will encounter. If the number is higher than approximately 50, the computation times to run the conformance algorithms become large fast, and it may lead to too many unique names, resulting in process models that have little additional value. Naming conventions have been mentioned as the most pressing issue for adequate process mining and this study confirms this.
- When the additional value of applying the process mining pipeline to other datasets is established, there can be made efforts to standardize the code-base. In this project, we observed that a lot of functionality in the pipeline is widely applicable, which can diminish the time needed to do a similar project. Then, more effort can be made into the visualizations, especially on the cost distributions and interactive Petri nets.
- Lastly, adequate data infrastructure is needed when using larger real-world datasets. For insightful process mining, rich event logs are essential. When using linked data from several sources, containing millions of records, the memory limits of reading directly from regular data files as CSV and converting to event logs can be reached. In this project in several instances the regular packages in both R and Python had trouble with the size of the records and this was also seen in test runs with ProM.

Regarding the outcomes of *the case study*, we state the following recommendations:

- Research the unexpected result that the costs of stage B patients receiving chemotherapy is higher than those of stage C patients. This result might be reversed when a larger sample size is evaluated, and if not, it would be interesting to find out what the characteristics are of patients that incurrer higher cost in stage B. A hypothesis can be that stage B patients that have recurring or metatstatic cancers do not have their stage updated and actually incur most of their costs when they should be classified as stage D.
- Research primary care more extensively and re-apply the methodology. The absolute number of that were linked to the primary care dataset and eventually included based on their sympotoms was relatively low (187 compared to the 3233 in hospital care), which could reduce the validity of the obtained models. As well, the process models obtained in this study could be improved by implementing a better suited classification scheme or implementation of a Natural Language Processing component in the workflow, to cluster groups of activities that are relatively the same together under a single name. This would yield better interpretable models, as well more valid models that describe the actual provided care.

# 8 CONCLUSION

This thesis set out to demonstrate process mining techniques and apply these on a linked dataset of real-world de-identified colorectal cancer patient data as a proof-of-concept study. A process mining workflow has been designed, consisting of joining datasets and selecting a cohort of interest, applying a process discovery algorithm to derive care pathways in consequent phases of care and applying conformance algorithms to evaluate the quality of these models. Additionally, the workflow incorporated a custom algorithm for adding aggregated costs as a numerical attribute to resulting process models and calculated distributions of costs for each of the subpopulations within the cohort, allowing comparison between them to identify disparities.

This workflow was applied onto a cohort of colorectal cancer patients, treated in 3 hospital groups in the Melbourne metropolitan area, Western Health, Royal Melbourne Hospital, and the Peter MacCallum Cancer Centre. These patients were linked to the Victorian Admitted Episodes Dataset (VAED) containing hospital information, the General Practitioner's primary care database Medicine Insight (NPS) and the registry Treatment of Recurrent and Advanced Colorectal Cancer (TRACC). Costing was based on prices in the Medicare Benefits Schedule (MBS) for primary care and on prices in the Pharmaceutical Benefit Scheme (PBS) for medication. The costs for the hospital's care was calculated with the Weighted Inlier Equivalent Separation (WIES).

The resulting pathways, as well as quality metrics for the pathways and enhanced models showing are displayed in an interactive app. The conformance algorithms for alignment-based fitness, precision generalization and for simplicity of the models incorporated in the workflow, provide how well each model represents the patients in a certain subpopulation, which is displayed in barcharts. From this, we can conclude which subpopulations align the best to this model, which turned out to be patients that are more close to the middle age-groups (50-70 year old), patients with colon cancer first and rectal cancer second, male patients a little more then female patients, patients that are not from Aboriginal or Torres Strait Islander descent the most and Aboriginal people secondly, patients that live in Major cities or Inner regions and patients that have Stage C cancer more than respectively D, B and A.

A case study was performed to evaluate differences in care and costs of care between colon cancer patients in different stages. Hospital admissions contain the largest part of the costs (93.34% of total costs), and lower stages stage A and B have their distributions more skewed towards the lower end, while C and specifically D are more skewed towards the higher costs. The Chemotherapy phase has a clear distinction in costs distribution between the stages, while within the other phases, the distributions are not so different. Unexpectedly, stage C patients had less expensive chemotherapy then stage B & D patients We found that the average costs of care for Stage A is 25.007 dollars, for Stage B it is 21.025 dollars Stage C it is 11.227 dollars, for Stage D it is 23.295 dollars and for the Unknown Stage it is 9.887 dollars. Most of the costs incurred in this phase come from the MFOLFOX 6 regimen and that the large differences per stage in this phase are mainly due to the large differences between stage B & D patients and stage C patients receiving this treatment regimen. Additional research into the unexplained discrepancy for chemotherapy costs of stage C patients is desired.

Within this study, Process Mining proved to be a value-adding method for providing insights on differences between patient groups in complex care. This methodology is data-driven in comparison to consensus-based guidelines like Optimal Care Pathways, and displays actual provided care on a detailed level, including deviations that doctors routinely make to accommodate for patientcharacteristics and -preference. The field of Process Mining is expected to grow rapidly over the next years and to be applied in case studies in health services research and other domains within the healthcare sector.

Additional research should focus on the primary care, as in this study, the number of patients linked to the primary care dataset was relatively low. The absolute number of that were linked to the primary care dataset and eventually included based on their sympotoms was relatively low (187 compared to the 3233 in hospital care), which could reduce the validity of the obtained models. As well, the process models for primary care obtained in this study could be improved by implementing a better suited classification scheme or implementation of a Natural Language Processing component in the workflow, to cluster groups of activities that are relatively the same, together under a single name. This would yield better interpretable models, as well more valid models that describe the actual provided care.

# 9 REFERENCES


[1] A. L. Siu, "The health services researcher, multiple identities," *Health Serv. Res.*, vol. 37, no. 1, pp. 3–6, 2002.

[2] World Health Organization International Agency for Research on Cancer (IARC)., "GLOBOCAN 2018: estimated cancer incidence, mortality and prevalence worldwide in 2018," 2018.

[3] R. L. Siegel *et al.*, "Colorectal cancer statistics, 2020," *CA. Cancer J. Clin.*, vol. 70, no. 3, pp. 145–164, May 2020.

[4] T. Winslow, "Digestive system," *Terese Winslow LLC, Medical and Scientific Illustration*, 2019. [Online]. Available: https://www.teresewinslow.com/digestion/aoz6hbz112a4qi1oikf6z6vyzuu134. [Accessed: 27-Oct-2020].

[5] D. Quinn and L. Shannon, "The colon and rectum.," *Neonatal network : NN*, vol. 19, no. 6. pp. 48–52, 2000.

[6] "Cancer." [Online]. Available: http://mrhardinsclass.weebly.com/cancer.html. [Accessed: 09-Mar-2021].

[7] Nci and Seer, "Colon, Rectosigmoid, and Rectum Equivalent Terms and Definitions C180-C189, C199, C209 (Excludes lymphoma and leukemia M9590-M9992 and Kaposi sarcoma M9140)."

[8] S. Y. Pan, "Epidemiology of cancer of the small intestine," *World J. Gastrointest. Oncol.*, vol. 3, no. 3, p. 1, 2011.

[9] "Bowel cancer (Colorectal cancer) in Australia statistics | Cancer Australia." [Online]. Available: https://www.canceraustralia.gov.au/affected-cancer/cancer-types/bowel-cancer/statistics. [Accessed: 15-Oct-2020].

[10] C. M. Johnson *et al.*, "Meta-Analyses of colorectal cancer risk factors," *Cancer Causes Control*, vol. 24, no. 6, pp. 1207–1222, Jun. 2013.

[11] P. Rawla, T. Sunkara, and A. Barsouk, "Epidemiology of colorectal cancer: Incidence, mortality, survival, and risk factors," *Przeglad Gastroenterologiczny*, vol. 14, no. 2. Termedia Publishing House Ltd., pp. 89–103, 2019.

[12] R. L. Siegel *et al.*, "Colorectal cancer statistics, 2020," *CA. Cancer J. Clin.*, vol. 70, no. 3, pp. 145–164, May 2020.

[13] AIHW, "Bowel cancer (colorectal cancer) in Australia," 2016.

[14] C. J. Parkin, S. W. Bell, and N. Mirbagheri, "Colorectal cancer screening in Australia: An update," *Aust. J. Gen. Pract.*, vol. 47, no. 12, pp. 859–863, Dec. 2018.

[15] S. Rasmussen, P. V Larsen, J. Søndergaard, S. Elnegaard, R. P. Svendsen, and D. E. Jarbøl, "Specific and non-specific symptoms of colorectal cancer and contact to general practice," *Fam. Pract.*, vol. 32, no. 4, p. cmv032, May 2015.

[16] N. Health and M. Research Council, "Clinical Practice Guidelines for the Prevention, Early Detection and Management of Colorectal Cancer."

[17] M. S. K. Thomas R, "Australian Clinical Practice Guidelines for the Prevention, Early Detection and Management of Colorectal Cancer.," *Natl. Heal. Med. Res. Counc. Aust. Gov.*, vol. 2, no. 2, pp. 38–39, 2020.

[18] Cancer Australia, "National Cancer Control Indicators," *Cancer Australia*, 2020. [Online]. Available: https://ncci.canceraustralia.gov.au/outcomes/cancer-mortality/cancer-mortality. [Accessed: 27-Oct-2020].

[19] A. Luck, C. Chow, C. Farmer, and H. Hicks, "Optimal approach to elective resection for colon cancers (COL1-2a) - Clinical Guidelines Wiki," *Clinical Guidelines Network*, 2017. [Online]. Available: https://wiki.cancer.org.au/australia/Clinical_question:Surgical_resection_colon_cancer. [Accessed: 28-Oct-2020].

[20] C. J. Allegra *et al.*, "Bevacizumab in stage II-III colon cancer: 5-year update of the National Surgical Adjuvant Breast and Bowel Project C-08 trial," *J. Clin. Oncol.*, vol. 31, no. 3, pp. 359–364, Jan. 2013.

[21] A. Luck, C. Chow, C. Farmer, and H. Hicks, "Optimal approach to elective resection for rectal cancers (REC1-2a) - Clinical Guidelines Wiki," *Clinical Guidelines Network*, 2017. [Online]. Available: https://wiki.cancer.org.au/australia/Clinical_question:Surgical_resection_rectal_cancer. [Accessed: 26-May-2021].

[22] S. Abbas, V. Lam, and M. Hollands, "Ten-Year Survival after Liver Resection for Colorectal Metastases: Systematic Review and Meta-Analysis," *ISRN Oncol.*, vol. 2011, pp. 1–11, Jun. 2011.

[23] S. Kopetz *et al.*, "Improved survival in metastatic colorectal cancer is associated with adoption of hepatic resection and improved chemotherapy," *J. Clin. Oncol.*, vol. 27, no. 22, pp. 3677–3683, Aug. 2009.

[24] S. Kopetz *et al.*, "Improved survival in metastatic colorectal cancer is associated with adoption of hepatic resection and improved chemotherapy," *J. Clin. Oncol.*, vol. 27, no. 22, pp. 3677–3683, Aug. 2009.

[25] T. X. Yang, B. Billah, D. L. Morris, and T. C. Chua, "Palliative resection of the primary tumour in patients with Stage IV colorectal cancer: Systematic review and meta-analysis of the early outcome after laparoscopic and open colectomy," *Colorectal Disease*, vol. 15, no. 8. Colorectal Dis, Aug-2013.

[26] M. E. Clark and R. R. Smith, "Liver-directed therapies in metastatic colorectal cancer," *Journal of Gastrointestinal Oncology*, vol. 5, no. 5. Pioneer Bioscience Publishing, pp. 374–387, 2014.

[27] G. Masi *et al.*, "Randomized trial of two induction chemotherapy regimens in metastatic colorectal cancer: An updated analysis," *J. Natl. Cancer Inst.*, vol. 103, no. 1, pp. 21–30, Jan. 2011.

[28] M. E. Porter, "A Strategy for Health Care Reform — Toward a Value-Based System," *N. Engl. J. Med.*, vol. 361, no. 2, pp. 109–112, Jul. 2009.

[29] E. Aviki, S. M. Schleicher, S. Mullangi, K. Matsoukas, and D. Korenstein, “Value-based healthcare delivery models in oncology: A systematic review.,” *J. Clin. Oncol.*, vol. 36, no. 15_suppl, p. 6525, May 2018.
[30] Cancer Council Australia, “Optimal care pathway for people with colorectal cancer,” 2016.
[31] A. R. C. Maita *et al.*, “A systematic mapping study of process mining,” *Enterprise Information Systems*, vol. 12, no. 5. Taylor and Francis Ltd., pp. 505–549, 28-May-2018.
[32] E. Rojas, J. Munoz-Gama, M. Sepúlveda, and D. Capurro, “Process mining in healthcare: A literature review,” *Journal of Biomedical Informatics*, vol. 61. Academic Press Inc., pp. 224–236, 01-Jun-2016.
[33] W. Van der Aalst, *Process mining: Data science in action*. 2016.
[34] T. Murata, “Petri Nets: Properties, Analysis and Applications,” *Proc. IEEE*, vol. 77, no. 4, pp. 541–580, 1989.
[35] M. De Leoni and W. M. P. Van Der Aalst, “Data-aware process mining: Discovering decisions in processes using alignments,” in *Proceedings of the ACM Symposium on Applied Computing*, 2013, pp. 1454–1461.
[36] G. Geleijnse *et al.*, “Using Process Mining to Evaluate Colon Cancer Guideline Adherence with Cancer Registry Data: a Case Study,” in *{AMIA} 2018, American Medical Informatics Association Annual Symposium, San Francisco, CA, November 3-7, 2018*, 2018.
[37] W. M. P. van der Aalst, “Process Design by Discovery: Harvesting Workflow Knowledge from Ad-hoc Executions,” in *Knowledge Management: An Interdisciplinary Approach*, 2000, pp. 9–10.
[38] R. Agrawal, D. Gunopulos, and F. Leymann, “Mining process models from workflow logs,” in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 1998, vol. 1377 LNCS, pp. 469–483.
[39] V. Verma *et al.*, “A systematic review of the cost and cost-effectiveness studies of immune checkpoint inhibitors 11 Medical and Health Sciences 1112 Oncology and Carcinogenesis,” *Journal for ImmunoTherapy of Cancer*, vol. 6, no. 1. BioMed Central Ltd., p. 128, 23-Nov-2018.
[40] A. J. M. M. Weijters and W. M. P. van der Aalst, “Process mining: Discovering workflow models from event-based data,” *Proc. 13th Belgium-Netherlands Conf. Artif. Intell.*, no. i, pp. 283–290, 2001.
[41] W. Van der Aalst, *Process mining: Data science in action*. Springer Berlin Heidelberg, 2016.
[42] L. Wen, J. Wang, W. M. P. Van Der Aalst, B. Huang, and J. Sun, “Mining process models with prime invisible tasks,” *Data Knowl. Eng.*, vol. 69, no. 10, pp. 999–1021, Oct. 2010.
[43] L. Wen, W. M. P. Van Der Aalst, J. Wang, and J. Sun, “Mining process models with non-free-choice constructs,” *Data Min. Knowl. Discov.*, vol. 15, no. 2, pp. 145–180, Oct. 2007.
[44] A. K. A. De Medeiros, B. F. Van Dongen, W. M. P. Van Der Aalst, and A. J. M. M. Weijters, “Process mining for ubiquitous mobile systems: An overview and a concrete algorithm,” *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)*, vol. 3272, pp. 151–165, 2004.
[45] A. J. M. M. Weijters, W. M. P. van der Aalst, and A. K. A. de Medeiros;, “Process Mining with the HeuristicsMiner Algorithm,” *Beta working papers*, no. May. 2006.
[46] W. M. P. Van Der Aalst, A. K. A. De Medeiros, and A. J. M. M. Weijters, “Genetic process mining,” in *Lecture Notes in Computer Science*, 2005, vol. 3536, pp. 48–69.
[47] N. Pelekis, B. Theodoulikis, I. Kopanakis, and Y. Theodoridis, “Fuzzy Miner: Extracting Fuzzy Rules from Numerical Patterns,” *Int. J. Data Warehous. Min.*, vol. 1, no. 1, pp. 57–81, Jan. 2005.
[48] 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 *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 2013, vol. 7927 LNCS, pp. 311–329.
[49] S. J. J. Leemans, D. Fahland, and W. M. P. van der Aalst, “Discovering block-structured process models from event logs containing infrequent behavior,” in *Lecture Notes in Business Information Processing*, 2014, vol. 171 171 LN, pp. 66–78.
[50] M. De Leoni, F. M. Maggi, and W. M. P. Van Der Aalst, “Aligning event logs and declarative process models for conformance checking,” in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 2012, vol. 7481 LNCS, pp. 82–97.
[51] F. Caron, J. Vanthienen, and B. Baesens, “Comprehensive rule-based compliance checking and risk management with process mining,” *Decis. Support Syst.*, vol. 54, no. 3, pp. 1357–1369, 2013.
[52] A. Rozinat and W. M. P. van der Aalst, “Conformance checking of processes based on monitoring real behavior,” *Inf. Syst.*, vol. 33, no. 1, pp. 64–95, Mar. 2008.
[53] A. Adriansyah, *Aligning observed and modeled behavior Aligning Observed and Modeled Behavior*. 2014.
[54] F. A. Yasmin, F. A. Bukhsh, and P. De Alencar Silva, “Process enhancement in process mining: A literature review,” in *CEUR Workshop Proceedings*, 2018, vol. 2270, pp. 65–72.
[55] W. E. Nauta, “Towards Cost-Awareness in Process Mining,” no. July, p. 62, 2011.
[56] D. Thabet, N. Ganouni, S. A. Ghannouchi, and H. H. Ben Ghezala, “Towards context-aware business process cost data analysis including the control-flow perspective: A process mining-based approach,” in *Advances in Intelligent Systems and Computing*, 2021, vol. 1181 AISC, pp. 193–204.
[57] A. Pika, M. T. Wynn, S. Budiono, A. H. M. Ter Hofstede, W. M. P. Van Der Aalst, and H. A. Reijers, “Towards Privacy-Preserving Process Mining in Healthcare.”
[58] D. Dakic, D. Stefanovic, I. Cosic, T. Lolic, and M. Medojevic, “Business process mining application: A literature review,” in *Annals of DAAAM and Proceedings of the International DAAAM Symposium*, 2018, vol. 29, no. 1, pp. 866–875.

[59] W. M. P. van der Aalst, “Process Mining Manifesto.”
[60] S. Yang *et al.*, “Process mining the trauma resuscitation patient cohorts,” in *Proceedings - 2018 IEEE International Conference on Healthcare Informatics, ICHI 2018*, 2018, pp. 29–35.
[61] W. Li, H. Zhu, W. Liu, D. Chen, J. Jiang, and Q. Jin, “An anti-noise process mining algorithm based on minimum spanning tree clustering,” *IEEE Access*, vol. 6, pp. 48756–48764, Aug. 2018.
[62] 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 *Lecture Notes in Computer Science*, 2005, vol. 3536, pp. 444–454.
[63] Gartner Peer Insights, “Process Mining Software Reviews 2021,” *Gartner Peer Insights*, 2021. [Online]. Available: https://www.gartner.com/reviews/market/process-mining. [Accessed: 10-Jun-2021].
[64] A. Berti, S. J. Van Zelst, W. M. P. Van Der Aalst, and F. Gesellschaf, “Process mining for python (PM4py): Bridging the gap between process- And data science,” in *CEUR Workshop Proceedings*, 2019, vol. 2374, pp. 13–16.
[65] G. Janssenswillen and B. Depaire, “BupaR: Business process analysis in R,” in *CEUR Workshop Proceedings*, 2017, vol. 1920.
[66] 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,” *Lect. Notes Comput. Sci.*, vol. 3536, pp. 444–454, 2005.
[67] R. S. Mans, M. H. Schonenberg, M. Song, W. M. P. Van Der Aalst, and P. J. M. Bakker, “Process mining in healthcare - A case study,” in *HEALTHINF 2008 - 1st International Conference on Health Informatics, Proceedings*, 2008, vol. 1, pp. 118–125.
[68] M. Ghasemi and D. Amyot, “Process mining in healthcare: A systematised literature review,” *Int. J. Electron. Healthc.*, vol. 9, no. 1, pp. 60–88, 2016.
[69] M. Binder *et al.*, “On analyzing process compliance in skin cancer treatment: An experience report from the evidence-based medical compliance cluster (EBMC 2),” in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 2012, vol. 7328 LNCS, pp. 398–413.
[70] P. Rattanavayakorn and W. Premchaiswadi, “Analysis of the social network miner (working together) of physicians,” in *International Conference on ICT and Knowledge Engineering*, 2015, vol. 2015-Decem, pp. 121–124.
[71] F. Caron, J. Vanthienen, K. Vanhaecht, E. Van Limbergen, J. Deweerdt, and B. Baesens, “A process mining-based investigation of adverse events in care processes,” *Heal. Inf. Manag. J.*, vol. 43, no. 1, pp. 16–25, 2014.
[72] D. M. V. Sato, S. C. De Freitas, M. R. Dallagassa, E. E. Scalabrin, E. A. P. Portela, and D. R. Carvalho, “Conformance checking with different levels of granularity : A case study on bariatric surgery,” *Proc. - 2020 13th Int. Congr. Image Signal Process. Biomed. Eng. Informatics, CISP-BMEI 2020*, pp. 820–826, Oct. 2020.
[73] R. S. Mans, W. M. P. van der Aalst, R. J. B. Vanwersch, and A. J. Moleman, “Process Mining in Healthcare: Data Challenges When Answering Frequently Posed Questions,” *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)*, vol. 7738 LNAI, pp. 140–153, 2012.
[74] M. Cho *et al.*, “Developing data-driven clinical pathways using electronic health records: The cases of total laparoscopic hysterectomy and rotator cuff tears,” *Int. J. Med. Inform.*, vol. 133, p. 104015, Jan. 2020.
[75] G. Ibanez-Sanchez *et al.*, “Toward Value-Based Healthcare through Interactive Process Mining in Emergency Rooms: The Stroke Case,” *Int. J. Environ. Res. Public Heal. 2019, Vol. 16, Page 1783*, vol. 16, no. 10, p. 1783, May 2019.
[76] T. Yampaka and P. Chongstitvatana, “An application of process mining for queueing system in health service,” *2016 13th Int. Jt. Conf. Comput. Sci. Softw. Eng. JCSSE 2016*, Nov. 2016.
[77] E. Helm, A. M. Lin, D. Baumgartner, A. C. Lin, and J. Küng, “Towards the use of standardized terms in clinical case studies for process mining in healthcare,” *Int. J. Environ. Res. Public Health*, vol. 17, no. 4, Feb. 2020.
[78] N. Martin *et al.*, “Recommendations for enhancing the usability and understandability of process mining in healthcare,” *Artif. Intell. Med.*, vol. 109, p. 101962, Sep. 2020.
[79] “Special Issue on Innovative informatics methods for process mining in health care,” 2021. [Online]. Available: https://www.journals.elsevier.com/journal-of-biomedical-informatics/call-for-papers/innovative-informatics-methods-for-process-mining-in-health. [Accessed: 09-Jul-2021].
[80] International Journal of Information Retrieval Research (IJIRR), “Calls for Papers (special): Recent Techniques and Trends of Information Retrieval Research,” 2021. [Online]. Available: https://www.igi-global.com/calls-for-papers-special/international-journal-information-retrieval-research/1178. [Accessed: 09-Jul-2021].
[81] G. Janssenswillen, B. Depaire, M. Swennen, M. Jans, and K. Vanhoof, “bupaR: Enabling reproducible business process analysis,” *Knowledge-Based Syst.*, vol. 163, pp. 927–930, 2019.
[82] “Alpha Algorithm - ML Wiki.” [Online]. Available: http://mlwiki.org/index.php/Alpha_Algorithm#Loops_of_Length_One. [Accessed: 13-May-2021].
[83] “BioGrid Australia - Cancer Data Collection Software.” [Online]. Available: https://www.biogrid.org.au/page/110/cancer-data-collection-software. [Accessed: 29-Dec-2020].
[84] “BioGrid Australia - BioGrid TRACC Clinical Registry.” [Online]. Available: https://www.biogrid.org.au/page/116/biogrid-tracc-clinical-registry. [Accessed: 04-Jan-2021].

[85] "Victorian Admitted Episodes Dataset - health.vic." [Online]. Available: https://www2.health.vic.gov.au/hospitals-and-health-services/data-reporting/health-data-standards-systems/data-collections/vaed. [Accessed: 04-Jan-2021].
[86] D. Busingye *et al.*, "Data Resource Profile: MedicineInsight, an Australian national primary health care database," *Int. J. Epidemiol.*, vol. 48, no. 6, pp. 1741-1741H, Jul. 2019.
[87] Australian Government: Department of Health, "MBS Online," *Medicare benefits schedule*, 2016. [Online]. Available: http://www9.health.gov.au/mbs/search.cfm?q=&Submit=&sopt=I. [Accessed: 11-Jun-2021].
[88] Australian Government Department of Health, "Pharmaceutical Benefits Scheme (PBS) | PBS Statistics," 2020. [Online]. Available: https://www.pbs.gov.au/info/browse/statistics. [Accessed: 11-Jun-2021].
[89] "National Efficient Price Determination | IHPA." [Online]. Available: https://www.ihpa.gov.au/what-we-do/national-efficient-price-determination. [Accessed: 10-Jun-2021].
[90] Australian Bureau for Statistics, "Consumer Price Index, Australia," 2018.
[91] A. Partington, M. Wynn, S. Suriadi, C. Ouyang, and J. Karnon, "Process mining for clinical processes: A comparative analysis of four australian hospitals," *ACM Trans. Manag. Inf. Syst.*, vol. 5, no. 4, Jan. 2015.
[92] J. Karnon and H. Haji Ali Afzali, "When to use Discrete Event Simulation (DES) for the economic evaluation of health technologies? A review and critique of the costs and benefits of DES," *PharmacoEconomics*, vol. 32, no. 6. Springer International Publishing, pp. 547–558, 2014.
[93] W. M. P. van der Aalst, "Process mining and simulation: A match made in heaven!," in *Simulation Series*, 2018, vol. 50, no. 10, pp. 39–50.
[94] S. Relijveld, M. Van den Ven, and H. Koffijberg, "Analysis of variation in clinical practice over time using process mining," *J. Biomed. Inform.*, no. Special Issue: Process Mining in Healthcare, 2021.

# 10 APPENDICES

## *Appendix A. TNM-staging description*

| T — primary tumour | |
|---|---|
| TX | Primary tumour cannot be assessed |
| T0 | No evidence of primary tumour |
| Tis | Carcinoma in situ: intramucosal (involvement of lamina propria with no extension through muscularis mucosae) |
| T1 | Tumour invades submucosa (through muscularis mucosae but not into the muscularis propria) |
| T2 | Tumour invades muscularis propria |
| T3 | Tumour invades through muscularis propria into pericolorectalic (subserosal) tissues |
| T4 | Tumour invades the visceral peritoneum or invades or adheres to adjacent organ or structure |
| T4a | Tumour penetrates to the surface of the visceral peritoneum (including gross perforation of the bowel through areas of inflammation to the surface of the visceral peritoneum) |
| T4b | Tumour directly invades or adheres to other organs or structures |
| N - regional lymph node | |
| NX | Regional lymph nodes cannot be assessed |
| NO | No regional lymph nodes metastases |
| N1 | One to three regional nodes are positive (tumour in lymph nodes measuring >0.2mm), or any number of tumour deposits are present and all identifiable lymph nodes are negative |
| N1a | One regional lymph node is positive |
| N1b | Two or three regional lymph nodes are positive |
| N1c | No regional lymph nodes are positive, but there are tumour deposits in the<br>• subserosa<br>• mesentery<br>• or non-peritonised pericolic or perirectal/mesorectal tissues |
| N2 | Four or more regional lymph nodes are positive |
| N2a | Four to six regional lymph nodes are positive |
| N2b | Seven or more regional lymph nodes are positive |
| M — distant metastasis | |
| MO | No distant metastasis by imaging, etc; no evidence of tumour in distant sites or organs (This category is not assigned by pathologists.) |
| M1 | Metastasis to one or more distant sites or organs or peritoneal metastasis is identified |
| M1a | Metastasis to one site or organ is identified without peritoneal metastasis |
| M1b | Metastases to two or more sites or organs is identified without peritoneal metastasis |
| M1c | Metastasis to the peritoneal surface is identified alone or with other site or organ metastases |

# *Appendix B. Linkage maps*

General Note: USI's can be coupled with a site-unique identifier (the UNIVID). The episode datafile contains both a patient identifier and an identifier for a registered episode in each of these sites. These episode-id's are also used in the other data tables in ACCORD as identifier for an episode and can be linked to both an overview of the treatment (TS_id or `Treatment Summary id`) as well as an overview of the medication linked to these episodes (CHEMOTREATMENTID and MEDICATIONID)

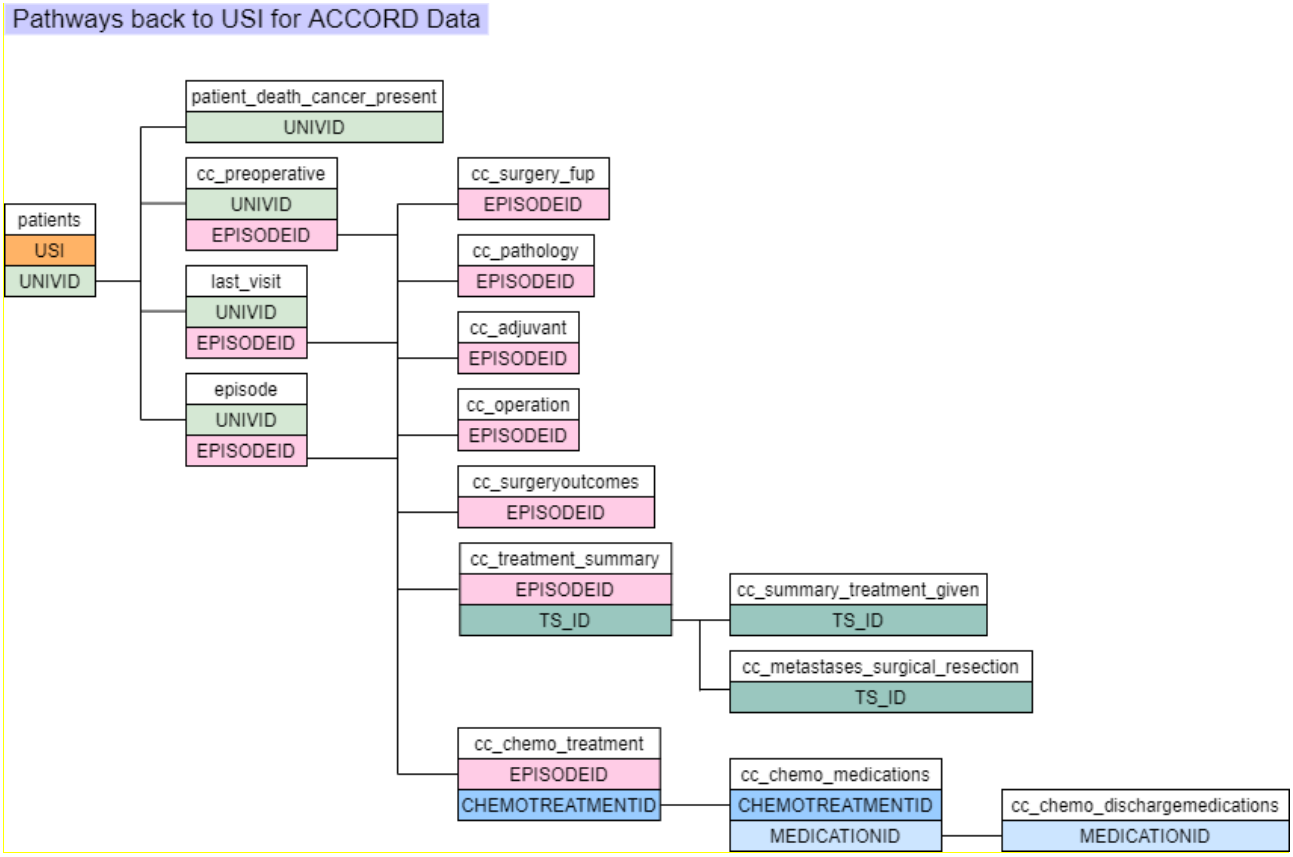



***Figure 57: Linkage pathway of ACCORD data tables***

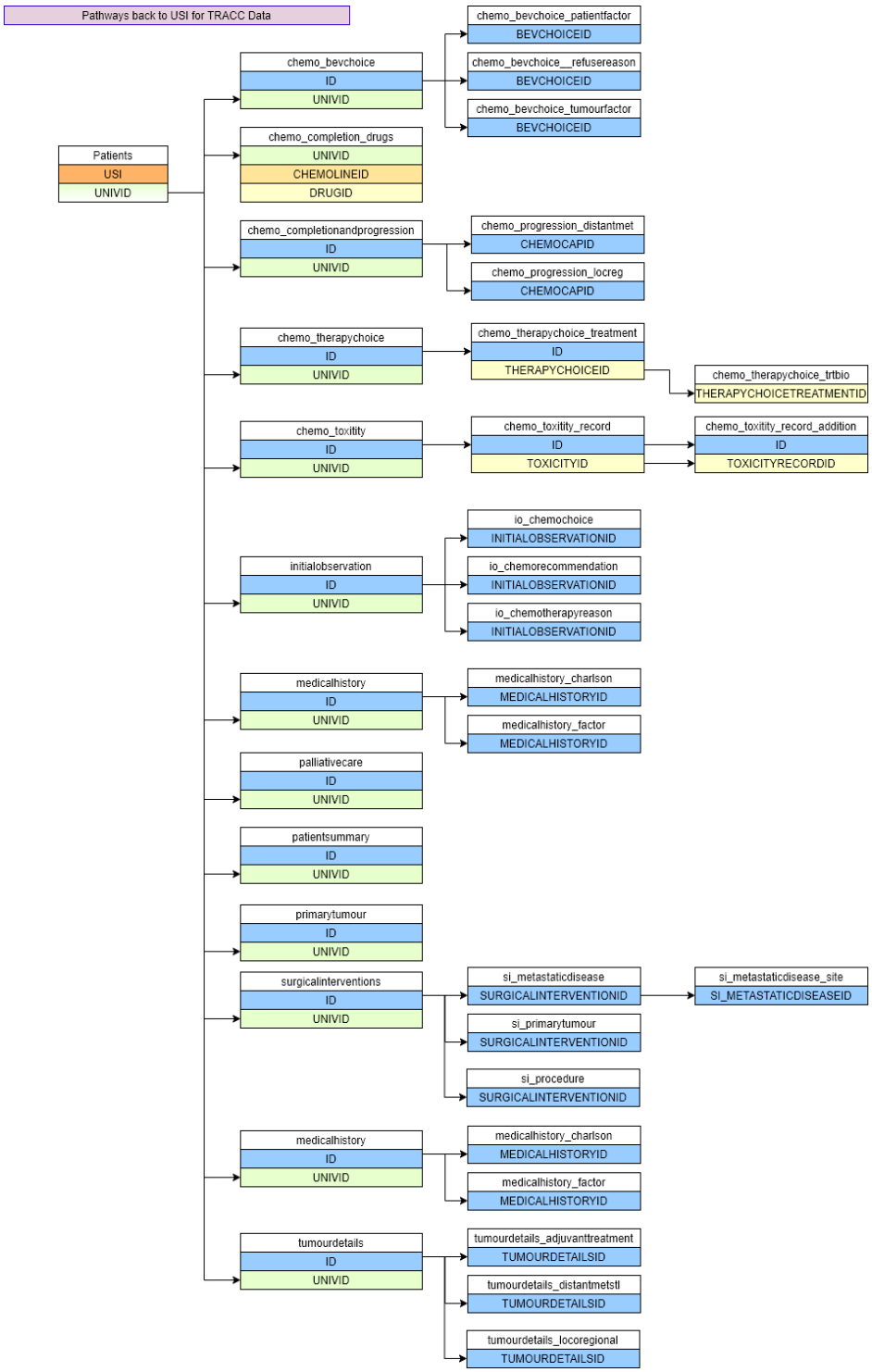



***Figure 58: Linkage pathway of TRACC data tables***

# *Appendix C. Statistics on patient characteristic distribution*

***Table 10: Statistics on patient characteristic distribution per dataset***

| | ACCORD LIFE EVENTS (N=4246) | ADMITTED EPISODES (N=3233) | CHEMO EPISODES (N=461) | DIAGNOSTIC TESTS (N=50) | GP VISITS (N=163) | PRESCRIPTIONS (N=84) |
|---|---|---|---|---|---|---|
| **Gender** | | | | | | |
| **F** | 1792 (42.2%) | 1357 (42.0%) | 175 (48.9%) | 24 (48.0%) | 74 (45.4%) | 43 (51.2%) |
| **M** | 2454 (57.8%) | 1876 (58.0%) | 183 (51.1%) | 26 (52.0%) | 89 (54.6%) | 41 (48.8%) |
| **Age Group** | | | | | | |
| **<30** | 42 (1.0%) | 38 (1.2%) | 7 (2.0%) | 2 (4.0%) | 2 (1.2%) | 2 (2.4%) |
| **30-39** | 113 (2.7%) | 84 (2.6%) | 16 (4.5%) | 1 (2.0%) | 6 (3.7%) | 2 (2.4%) |
| **40-49** | 311 (7.3%) | 247 (7.6%) | 40 (11.2%) | 4 (8.0%) | 18 (11.0%) | 7 (8.3%) |
| **50-59** | 698 (16.4%) | 522 (16.1%) | 82 (22.9%) | 8 (16.0%) | 29 (17.8%) | 9 (10.7%) |
| **60-69** | 1220 (28.7%) | 950 (29.4%) | 110 (30.7%) | 16 (32.0%) | 55 (33.7%) | 32 (38.1%) |
| **70-79** | 1181 (27.8%) | 902 (27.9%) | 70 (19.6%) | 11 (22.0%) | 30 (18.4%) | 19 (22.6%) |
| **80-89** | 572 (13.5%) | 416 (12.9%) | 29 (8.1%) | 8 (16.0%) | 20 (12.3%) | 10 (11.9%) |
| **90+** | 37 (0.9%) | 27 (0.8%) | 1 (0.3%) | 0 (0.0%) | 1 (0.6%) | 2 (2.4%) |
| **Unknown AGE** | 72 (1.7%) | 47 (1.5%) | 3 (0.8%) | 0 (0.0%) | 2 (1.2%) | 1 (1.2%) |
| **Tumour location** | | | | | | |
| **Colon** | 2580 (60.8%) | 1983 (61.3%) | 218 (60.9%) | 31 (62.0%) | 95 (58.3%) | 52 (61.9%) |
| **Rectal** | 1508 (35.5%) | 1153 (35.7%) | 132 (36.9%) | 19 (38.0%) | 64 (39.3%) | 30 (35.7%) |
| **Other** | 19 (0.4%) | 17 (0.5%) | 2 (0.6%) | 0 (0.0%) | 0 (0.0%) | 0 (0.0%) |
| **Undefined** | 139 (3.3%) | 80 (2.5%) | 6 (1.7%) | 0 (0.0%) | 4 (2.5%) | 2 (2.4%) |
| **Tumour Stage** | | | | | | |
| **A** | 763 (18.0%) | 591 (18.3%) | 37 (10.3%) | 10 (20.0%) | 37 (22.7%) | 17 (20.2%) |
| **B** | 1250 (29.4%) | 923 (28.5%) | 77 (21.5%) | 17 (34.0%) | 43 (26.4%) | 18 (21.4%) |
| **C** | 1037 (24.4%) | 802 (24.8%) | 111 (31.0%) | 10 (20.0%) | 30 (18.4%) | 19 (22.6%) |
| **D** | 646 (15.2%) | 526 (16.3%) | 106 (29.6%) | 9 (18.0%) | 34 (20.9%) | 25 (29.8%) |
| **Unknown Stage** | 550 (13.0%) | 391 (12.1%) | 27 (7.5%) | 4 (8.0%) | 19 (11.7%) | 5 (6.0%) |
| **Etnicity/ Indigenous Status** | | | | | | |
| **Aboriginal** | 507 (12.0%) | 297 (9.2%) | 18 (3.9%) | 10 (20.0%) | 18 (11.0%) | 9 (10.7%) |
| **Not Ab/TS** | 3462 (81.6%) | 2824 (87.4%) | 335 (72.7%) | 36 (72.0%) | 135 (82.8%) | 70 (83.3%) |
| **Torres Strait** | 18 (0.4%) | 18 (0.6%) | 1 (0.2%) | 0 (0.0%) | 0 (0.0%) | 0 (0.0%) |
| **Unknown Etnicity** | 254 (6.0%) | 93 (2.9%) | 107 (23.2%) | 4 (8.0%) | 10 (6.1%) | 5 (6.0%) |
| **Remoteness** | | | | | | |
| **Inner Regional** | 224 (5.3%) | 151 (4.7%) | 18 (5.0%) | 0 (0.0%) | 0 (0.0%) | 1 (1.2%) |
| **Major City** | 3959 (93.2%) | 3056 (94.5%) | 338 (94.4%) | 48 (96.0%) | 160 (98.2%) | 82 (97.6%) |
| **Outer Regional** | 36 (0.8%) | 18 (0.6%) | 0 (0.0%) | 0 (0.0%) | 1 (0.6%) | 0 (0.0%) |
| **Remote** | 1 (0.0%) | 0 (0.0%) | 0 (0.0%) | 0 (0.0%) | 0 (0.0%) | 0 (0.0%) |
| **Unknown** | 26 (0.6%) | 8 (0.2%) | 2 (0.6%) | 2 (4.0%) | 2 (1.2%) | 1 (1.2%) |

# *Appendix D. Symptoms included in GP visits*

***Table 11: Symptoms included in GP visits***

| ***Included symptoms*** | ***Alternative writing/ REGEX*** |
|---|---|
| Bowel cancer | |
| Cancer Pain | |
| Abdominal pain | (lower) Abdo pain |
| Diarrhoea | agrepl(, max.dist=2 |
| Palliative care | [Palliat] |
| Weight Loss | [Weight] |
| Rectal | [Rectal] |
| Colon | [Colon] |
| Dyspepsia | agrepl(, max.dist=2 |
| Occult blood in faeces | Occult blood in faeces – test for |
| Anaemia | agrepl(, max.dist=2 |
| Metastasis liver | [Metastasis] |

***source: D. Quinn and L. Shannon, "The colon and rectum.," Neonatal network : NN, vol. 19, no. 6. pp. 48–52, 2000 [5]***

# *Appendix E. CPI Price indices and NEP/NWAU values*

***Table 12: Quarterly and average Consumer price Index (CPI%) for Health***

| | year | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Quarters | 2006 | 2007 | 2008 | 2009 | 2010 | 2011 | 2012 | 2013 | 2014 | 2015 | 2016 | 2017 | 2018 | 2019 | 2020 |
| dec | -0,8 | -1,0 | -1,2 | -0,9 | -1,2 | -1,2 | -0,9 | -0,5 | -0,9 | -0,4 | -0,6 | -0,5 | -0,4 | -0,3 | 1,3 |
| sep | -0,7 | -0,5 | -0,2 | -1,0 | -0,7 | -1,0 | 2,4 | 0,0 | -0,2 | 0,3 | -0,2 | -0,2 | -0,4 | -0,2 | -0,1 |
| jun | 2,4 | 2,1 | 2,4 | 2,3 | 2,2 | 2,0 | 1,5 | 1,9 | 2,9 | 2,7 | 2,6 | 2,7 | 1,9 | 1,8 | -0,2 |
| mar | 4,4 | 3,5 | 4,0 | 4,4 | 4,7 | 3,9 | 4,4 | 3,0 | 2,6 | 2,5 | 1,9 | 2,0 | 2,2 | 1,9 | 1,7 |
| average CPI per year | 1,325 | 1,025 | 1,250 | 1,200 | 1,250 | 0,925 | 1,850 | 1,100 | 1,100 | 1,275 | 0,925 | 1,000 | 0,825 | 0,800 | 0,675 |

***Table 13: NEP/NWAU values for each year of interest in the dataset***

| | Year | NEP/NWAU | CPI% till 2020 | CPI (%) |
|---|---|---|---|---|
| NWAU | 2006-07 | $3.659 | 15,200 | 0,1520 |
| | 2007-08 | $3.804 | 14,175 | 0,1418 |
| | 2008-09 | $4.018 | 12,925 | 0,1293 |
| | 2009-10 | $4.307 | 11,725 | 0,1173 |
| | 2010-11 | $4.395 | 10,475 | 0,1048 |
| | 2011-12 | $4.544 | 9,550 | 0,0955 |
| NEP | 2012-13 | $4.808 | 7,700 | 0,0770 |
| | 2013-14 | $4.993 | 6,600 | 0,0660 |
| | 2014-15 | $5.007 | 5,500 | 0,0550 |
| | 2015-16 | $4.971 | 4,225 | 0,0423 |
| | 2016-17 | $4.883 | 3,300 | 0,0330 |
| | 2017-18 | $4.910 | 2,300 | 0,0230 |
| | 2018-19 | $5.012 | 1,475 | 0,0148 |
| | 2019-20 | $5.134 | 0,675 | 0,0068 |

# *Appendix F. Computation Times*

***Table 14: Computation times of Discovery & Conformance pipeline with various parameters.***

| Dataset | Runtype | Parameters | Duration |
|---|---|---|---|
| Admitted Episodes | Complete complement run | {inductive miner, imf, ALLGROUP, ALLVALUES} | 25h 46m (92.783 s) |
| Chemo Episodes | Complete complement run | {inductive miner, imf, ALLGROUP, ALLVALUES} | 36m |
| Prescriptions | Complete complement run | {inductive miner, imf, ALLGROUP, ALLVALUES} | 5m50s |
| GP Encounters | Complete complement run | {inductive miner, imf, ALLGROUP, ALLVALUES} | 11m23s |
| Diagnostic Tests | Complete complement run | {inductive miner, imf, ALLGROUP, ALLVALUES} | 4m45s |
| Admitted Episodes | Comparator CRC_TYPE “Colon”, ALL stages | {inductive miner, imf, CRCTYPE: Colon, STAGING_ACPS: ALLVALUES} | 18m30 |
| Chemo Episodes | Comparator CRC_TYPE “Colon”, ALL stages | {inductive miner, imf, CRCTYPE: Colon, STAGING_ACPS: ALLVALUES} | 2 minutes |
| Prescriptions | Comparator CRC_TYPE “Colon”, ALL stages | {inductive miner, imf, CRCTYPE: Colon, STAGING_ACPS: ALLVALUES} | 124 seconds |
| GP Encounters | Comparator CRC_TYPE “Colon”, ALL stages | {inductive miner, imf, CRCTYPE: Colon, STAGING_ACPS: ALLVALUES} | 136 seconds |
| Diagnostic Tests | Comparator CRC_TYPE “Colon”, ALL stages | {inductive miner, imf, CRCTYPE: Colon, STAGING_ACPS: ALLVALUES} | 111 seconds |
| Entire Pathway | Comparator CRC_TYPE “Colon”, ALL stages | {inductive miner, imf, CRCTYPE: Colon, STAGING_ACPS: ALLVALUES} | 23 minutes |

# *Appendix G. Frequency-annotated main pathways per phase*

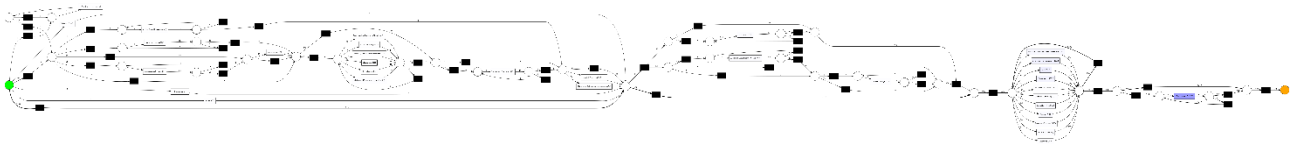


***Figure 59: Resulting pathway Admitted Episodes***

***Figure 60: Resulting pathway of GP Visits***

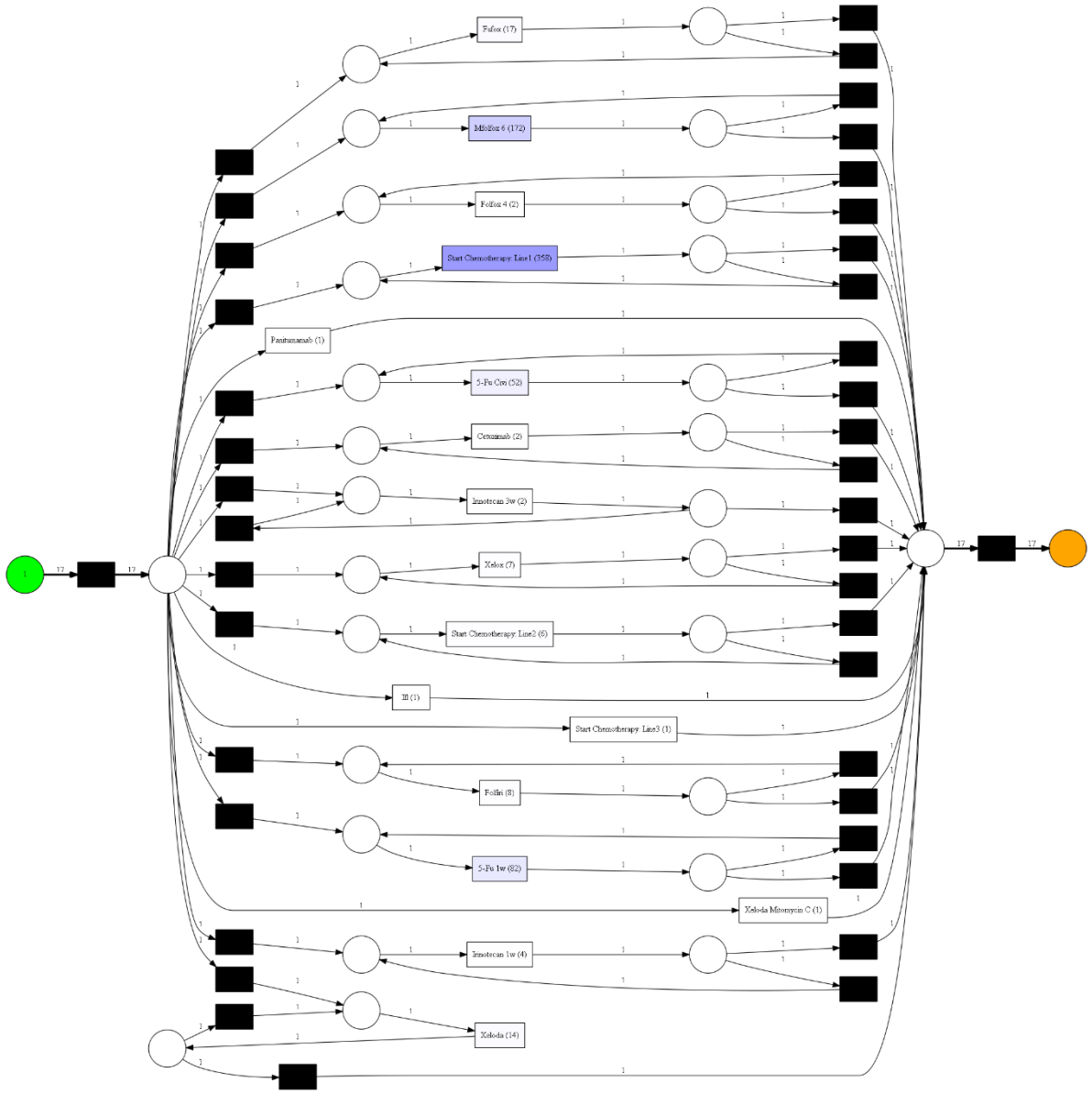



***Figure 61: Resulting pathway of Chemotherapy.***

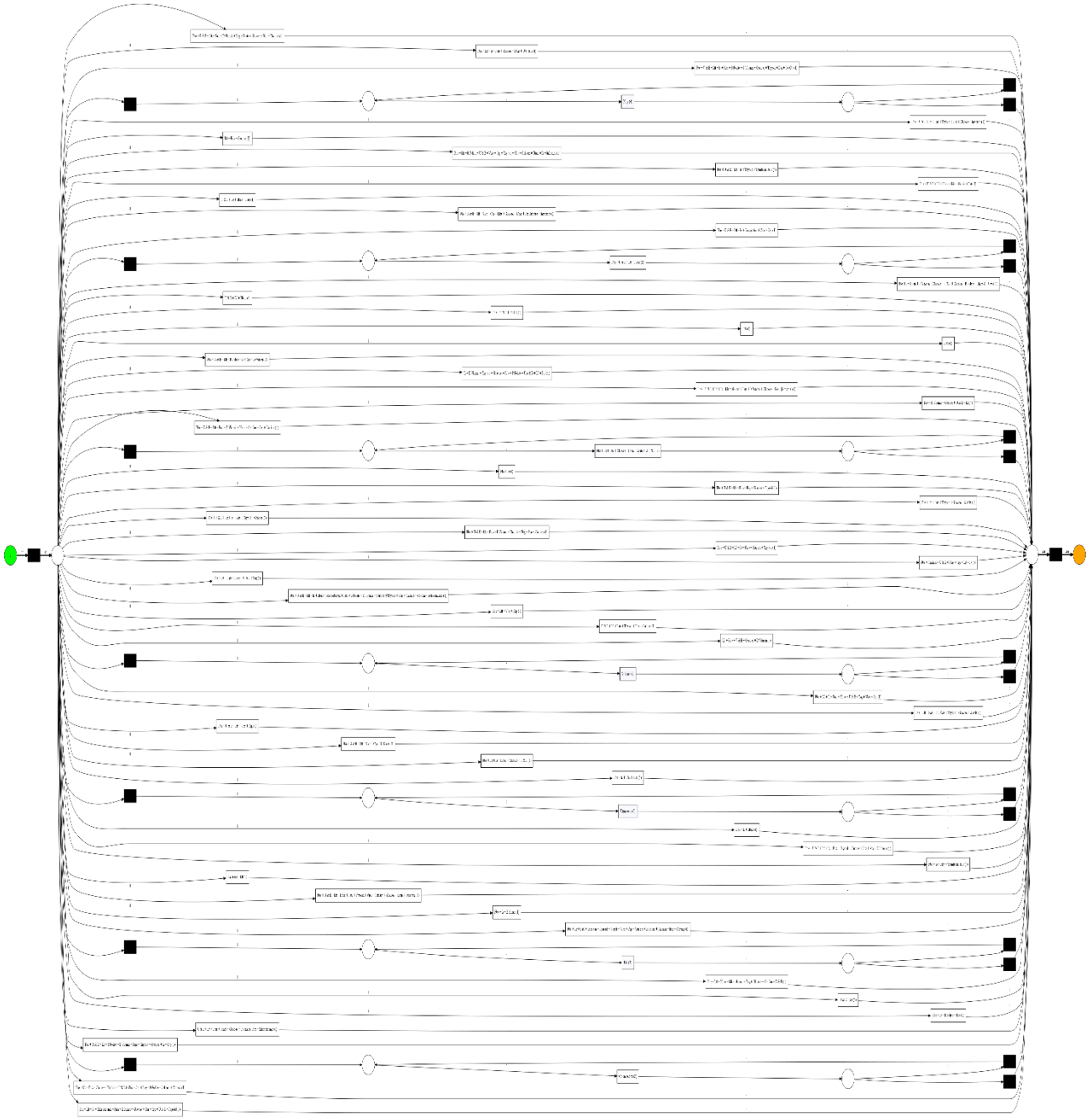


***Figure 62: Resulting pathway of Diagnostic tests.***

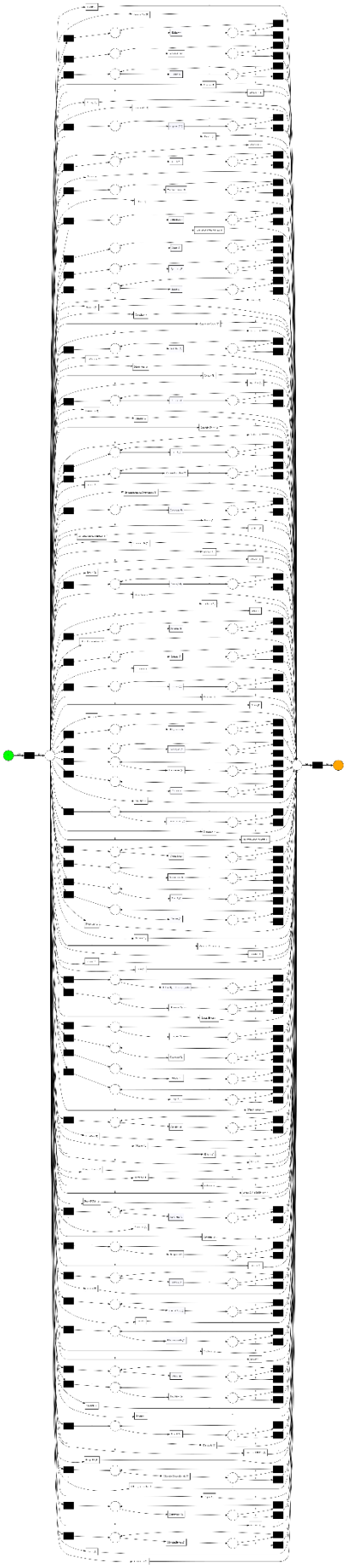
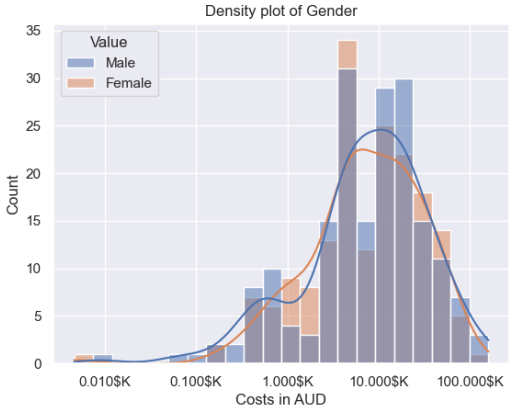
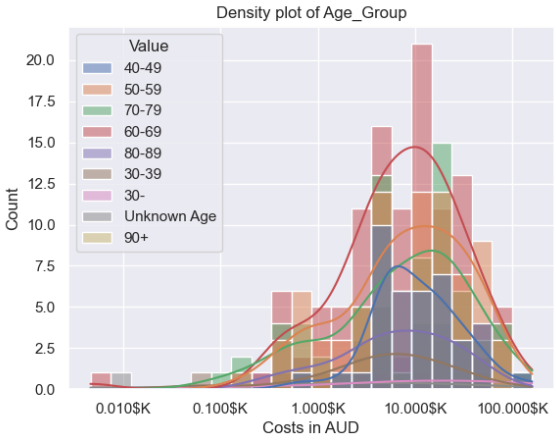


*Figure 63: Resulting pathway of Prescriptions.*

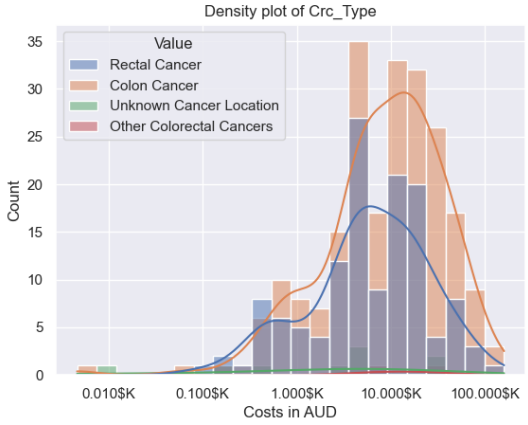
# *Appendix H. Chemotherapy cost distribution per characteristic*



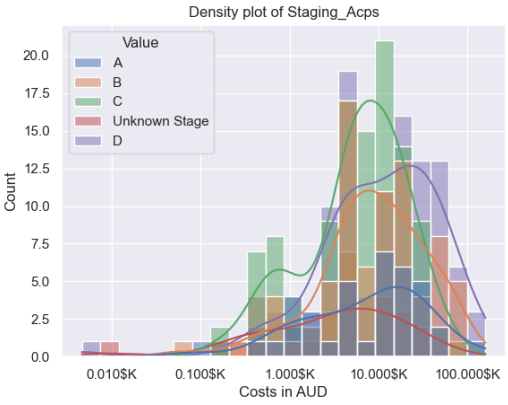

a



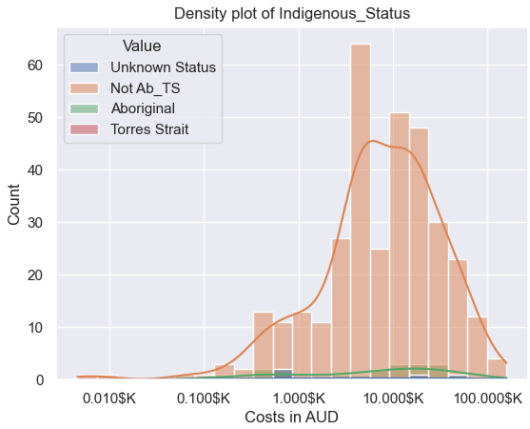

b



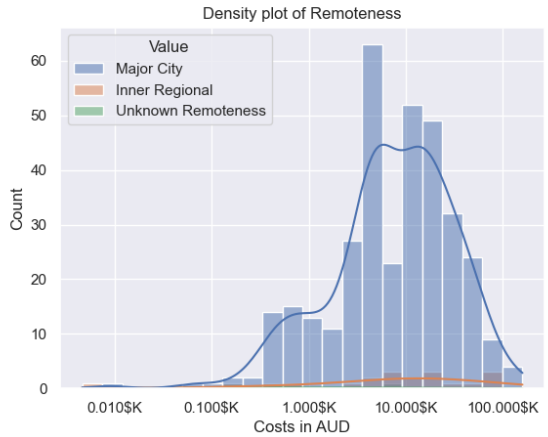

c




d




e




f

***Figure 64a-f: Density plots of total costs of the 'Chemo episodes' dataset***

# *Appendix I. Total pathway Colon Cancer*

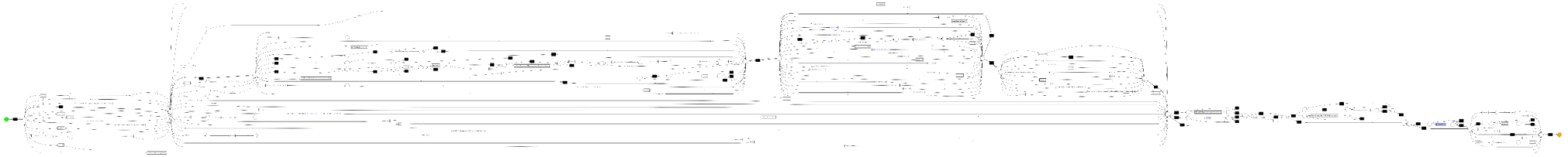


***Figure 65: Entire integrated pathway of Colon Cancer, annotated with Frequency.***