Processing Mining and Epidemiology: understanding patient journeys using care records



Lara Sophia Camille Chammas

St Anne's College

University of Oxford

A thesis submitted for the degree of

Doctor of Philosophy

Michaelmas 2024/2025

Abstract

Cancer services within the English National Health Service (NHS) are under pressure. Waiting times for everything from diagnosis to treatment are increasing and delivery targets are not being met. There is a need for reliable, comprehensible information and analysis techniques to support service monitoring, evaluation, and redesign. A greater degree of automation is needed to realise this potential: in particular, in the analysis of patient journeys to determine whether care is being delivered as intended and to understand how patterns of care affect outcomes.

Process mining techniques can be used to address exactly this need. However, such techniques can be difficult to deploy in a healthcare context due to a lack of appropriate methodology. This thesis argues for closer alignment with epidemiology and other clinical research methods to address the challenges and limitations of process mining for healthcare. The result is a methodology that can be used to mine patient journeys from electronic health records data that is routinely collected and made available.

A first method is an epidemiology-informed approach to process mining. It extends an existing method with additional steps from epidemiology and clinical research to produce a generalisable method for mining patient journeys. A second method involves the automation of the first stage of process mining, event-log generation, through the extraction of code-lists from clinical ontologies. A third involves the application of epidemiological methods in the analysis stage of process mining: in effect, process-informed epidemiology.

This methodology is demonstrated using actual patient data: in the analysis of treatment variation in colon cancer delivery at a major hospital, and in the mapping of patient journeys for late-stage rectal cancer patients across the country. Data availability, however, limits the granularity of discovered journeys, prevents conformance checking with national standards, and limits clinical conclusions that can be drawn. Discovered processes currently at most can highlight what is happening in care deliver, but cannot yet be used for evaluation of whether care is optimal or to inform new clinical guidelines. Throughout the thesis the need for increased data collection, specifically on intended care plans and clinical decision making, is argued as a requirement to support process mining techniques to analyse patient journeys. The analyses serve to illustrate the argument of the thesis: that an appropriate combination of epidemiology and process mining can deliver useful, actionable insights.

Contents

1 Introduction **1**

1.1 Motivation 1
1.2 Contribution 3
1.3 Thesis Structure 4

2 Background **6**

2.1 Introduction 6
2.2 Delivery of Care Using Pathways 7
2.3 The Patient Care Record 10
2.3.1 Clinical Coding 11
2.4 Learning Health Systems 13
2.5 Process Mining 14
2.5.1 Event-log Generation 15
2.5.2 Process Discovery 16
2.5.3 Conformance Checking 17
2.5.4 Process Enhancement 18
2.6 Process Mining for Healthcare (PM4H) 18
2.7 Process Mining and Clinical Research 20
2.8 Colorectal Cancer Care in the English NHS 21
2.9 English Cancer Care Datasets 22

3 An Epidemiology-informed Process Mining Approach **25**

3.1 Introduction 25
3.2 The Approach 28
3.2.1 Defining a Research Question 29
3.2.2 Literature Review 30
3.2.3 Patient Population and Subcohorts 31
3.2.4 Event-log Generation 32
3.2.5 Mining for Processes and Comparison of Results 33
3.2.6 Review Results and Iterate with Domain Experts 33
3.3 Colon Cancer Treatment Journeys at a NHS Trust 34
3.3.1 Research Question and Literature Review 34

3.3.2 Patient Population and Subcohorts . . . 35
3.3.3 Event Log Generation . . . 37
3.3.4 Results . . . 37
3.3.4.1 Population Level View . . . 38
3.3.4.2 Age . . . 40
3.3.4.3 Charlson Score . . . 43
3.3.4.4 Socio-economic Factors . . . 45
3.3.4.5 Time to Major Resection . . . 48
3.4 Importance of Data Provenance . . . 49
3.5 Discussion . . . 50

4 Which Events are in a Patient's Journey? 53
4.1 Introduction . . . 53
4.2 Preparing Healthcare Data . . . 55
4.2.1 Event Selection in PM4H . . . 56
4.2.2 Code-lists in Clinical Research . . . 57
4.3 Ontology Based Code-list Generation . . . 58
4.4 Study 1 . . . 60
4.4.1 Queries . . . 60
4.4.1.1 Recall and Precision . . . 61
4.4.2 Activity Abstraction . . . 63
4.4.2.1 Methods . . . 64
4.4.2.2 Method Comparison . . . 65
4.4.2.3 Effects on Trace Variants . . . 66
4.5 Study 2 . . . 69
4.5.1 Recall and Precision of Query Code-lists . . . 71
4.5.2 Discovered Processes . . . 73
4.5.2.1 Colon cancer . . . 73
4.5.2.2 Lung Cancer . . . 75
4.5.3 Enriching Code-lists with Statistics . . . 77
4.6 Discussion . . . 79

5 Understanding Stage IV Rectal Cancer Care Pathways 82
5.1 Introduction . . . 82
5.2 Analysing Rectal Cancer . . . 83
5.3 Methods . . . 85
5.3.1 Defining the Patient Population and Subcohorts . . . 85
5.3.2 Identifying Events of Interest . . . 85
5.3.3 Event Log Generation and Patient Journey Classification . . . 88
5.3.4 Analysis . . . 88

5.4 Results . . . 89
5.4.1 Patient Demographics and Journey Breakdowns . . . 89
5.4.2 Patient Journeys . . . 91
5.4.2.1 Events Along the Journey . . . 93
5.4.3 Factors Affecting Treatment Completion . . . 97
5.4.4 Treatment Pathways by Cancer Alliance and MDT . . . 98
5.4.5 Effects of Complete Pathway Order . . . 103
5.4.5.1 Variation by Cancer Alliance and MDT . . . 103
5.4.5.2 Survival . . . 104
5.5 Discussion . . . 112

6 Thesis Discussion **114**
6.1 Summary of Findings . . . 114
6.2 Limitations . . . 116
6.3 Future Work . . . 117

Appendices

A Additional DFGs **120**

B Patient Demographics by Cancer Alliance **122**

C Patient Demographics by first SACT regimen **124**

D KM Curves for complete pathways starting with primary site directed therapy **126**

E KM Curves for complete pathways starting with metastasis directed therapy **129**

References **132**

The NHS will last as long as there are folk left with faith to fight for it

— Aenurin Bevan

1 Introduction

Contents

1.1 Motivation **1**
1.2 Contribution **3**
1.3 Thesis Structure **4**

1.1 Motivation

Cancer services within the English National Health Service (NHS) are under pressure. Waiting times for everything from diagnosis to treatment are increasing and delivery targets are not being met. Statistics on cancer outcomes suggest that the NHS is performing poorly in comparison to care systems in other countries and show that there are significant inequalities across the population [1], [2]. This situation is exacerbated by a lack of investment, and a reduction in capacity during and post the COVID-19 pandemic [3], [4].

Furthermore, the demand for cancer care services is increasing: our population is ageing, and older people are more likely to be diagnosed with cancer; also, more effective therapies are available and patients are surviving for longer. Returning to previous levels of investment and capacity will not be enough. Health systems must find ways to improve the efficiency, effectiveness, and equity of care, as

well as generate a feedback mechanism to disseminate clinical learnings back into guidelines [5].

To do this, accurate, detailed, and timely information on all aspects of care delivery is required. Health systems need reliable, comprehensive data on how patients are diagnosed, how they are treated, and whether or not this treatment is working. Once this feedback loop is closed, there will be more effective mechanisms for identifying needed changes to services delivery, and for monitoring the effect of introduced changes.

The ambition is that of a learning health system: a healthcare system capable of continuous improvement and innovation [6]–[8]. However, as Foley and Vale [5] observe, a number of 'technical building blocks' are required to support the virtuous cycle of practice-to-data, data-to-knowledge, and knowledge-to-practice.

This thesis argues that process mining, a set of analytical techniques for discovering and analysing the processes behind real-world event data, can provide one of these building blocks: supporting the transformation of data into knowledge [9], [10]. Process mining techniques are already being used to do this in other sectors, answering questions about variation, compliance, and efficiency [11].

The application of process mining to the healthcare context is a growing field of research, with various literature reviews written on its applications, including to oncology [12]–[14]. The growing consensus, however, is that it is still difficult to meaningfully discover processes from patient care records that are clinically relevant and can aid in improving care [10], [15].

Current epidemiology and clinical research on cancer care focuses on a particular treatment and studying where it is occurring, in which populations, and whether differences are associated with changes in outcomes [16], [17]. The focus is on a particular instance in a care pathway. Cancer treatments are not delivered individually however, but as an ongoing series of treatments which are planned around patient preference and evolving health state.

Therefore, this thesis proposes and explores approaches that integrate and align process mining techniques and epidemiology research methods to enhance

and expand both research fields. It then demonstrates the application of these approaches in the analysis of treatment variation and patient journeys for two common cancers, using data drawn from NHS patient records.

1.2 Contribution

This thesis proposes new methods that integrate process mining techniques and epidemiology research methods to make healthcare process mining results clinically significant and epidemiology research more holistic. These methods aim to derive patient journeys using routine electronic health record data in England, focusing on colorectal cancer patients.

The first method is an approach called epidemiology-informed process mining. It extends an established process mining method with steps taken from epidemiology and clinical research methodology, together with some alignment of terminology and an increased emphasis on iteration. It aims to increase the reproducibility of process mining insights, increase the clinical relevancy of these insights, and provide a flexible approach that generalises to any healthcare dataset.

The second method is for activity selection during event-log generation for the process mining for healthcare community. The method draws a parallel between activity selection, as performed in process mining, and code-list curation, as used in clinical research, by exploiting the domain knowledge stored within the SNOMED clinical ontology. The method investigates the extent to which event-log generation, including event abstraction, can be automated.

Alongside increasing the collaboration between the process mining and epidemiology communities, this thesis actively advocates for strengthening the understanding of the provenance of patient care record data used during healthcare research.

These methods are applied to two case-studies of healthcare research. The first is to determine if there is variation in colon cancer delivery between patient groups in a regional NHS Trust. Results showed that the Trust delivers similar levels of care between deprivation groups and varied care between age groups and

patient comorbidity scores, which aligns with documented national variation and the need for personalised care for different health states.

The second study is to understand stage IV rectal cancer patient journeys in England's NHS. Results showed that many treatment pathways are occurring nationally, which can be grouped into 8 un-ordered distinct treatment pathways and three treatment types: incomplete, partial, and complete treatment. Variation in delivery of treatment pathways and treatment types can be seen across colorectal cancer multi-disciplinary teams (MDTs, a group of care professionals from different clinical disciplines who come together to make decisions regarding complex patient treatment) and Cancer Alliances (partnerships between local health and social care organisations to improve local cancer care planning and delivery.). This variation goes down even to the order of treatments delivered within the complete treatment type pathways. Differences in survival are shown between different treatment orders, but the reason for this difference cannot be deduced due to insufficient data depth.

Overall, this thesis contributes to both the process mining and epidemiology research communities by demonstrating how the alignment of techniques can generate clinically meaningful insights that go beyond investigating specific instances of care to consider the entire patient pathway as delivered.

1.3 Thesis Structure

Chapter 2 of this thesis presents background information related to the development of this thesis. This includes outlining what is a patient pathway, how patient care records are generated, what is process mining, and how it has been applied to healthcare previously.

Chapter 3 presents the epidemiology-informed process mining approach and its application to understanding variation in colon cancer patient journeys at a major regional NHS Trust.

Chapter 4 presents a new method for healthcare activity selection to generate event-logs for process mining. It will also explore methods for event abstraction and the effects of code-lists and clinical terminologies on process mining results.

Chapter 5 utilises the methods in Chapters 3 and 4 to investigate treatment pathways for stage IV rectal cancer patients across England's NHS. It determines common treatment pathways and checks for variation across MDTs and Cancer Alliances. It also shows how different treatment pathways may be affecting survival outcomes in patients. This chapter focuses more on generating epidemiology outputs, and therefore the method is referred to as process-informed epidemiology.

Chapter 6 summarises the findings and contributions of each chapter and will discuss limitations of the thesis as a whole, as well as areas for future work.

"Measurement is the first step that leads to control and eventually to improvement. If you can't measure something, you can't understand it. If you can't understand, it you can't control it. If you can't control it, you can't improve it"

— H. James Harrington

2 Background

Contents

2.1 Introduction 6
2.2 Delivery of Care Using Pathways 7
2.3 The Patient Care Record 10
2.3.1 Clinical Coding 11
2.4 Learning Health Systems 13
2.5 Process Mining 14
2.5.1 Event-log Generation 15
2.5.2 Process Discovery 16
2.5.3 Conformance Checking 17
2.5.4 Process Enhancement 18
2.6 Process Mining for Healthcare (PM4H) 18
2.7 Process Mining and Clinical Research 20
2.8 Colorectal Cancer Care in the English NHS 21
2.9 English Cancer Care Datasets 22

2.1 Introduction

This chapter sets the stage for the thesis by providing a background introduction to the methods used and the context for the thesis findings. It describes how clinical and patient pathways can be defined, how patient care records are generated, and introduces the concepts of a learning health system and continuity of care. The field of process mining is then introduced, including its applications in healthcare

as a method for understanding patient pathways. The chapter concludes with an introduction to colorectal cancer and its treatment guidelines in England, including a description of the datasets used to understand cancer care in England.

2.2 Delivery of Care Using Pathways

One way healthcare systems can structure their processes to ensure consistent delivery of evidence-based care is through pathways [18]. A clinical, or care, pathway is a description of the programme of appointments, tests, treatments, and other actions recommended for a specific group of patients, such as those with a particular diagnosis or set of symptoms [19]–[22]. These pathways are meant to reflect up-to-date scientific understanding of diseases and symptoms, and the 'best practices' for care delivery within resource constraints. They provide a basis for ensuring continuity of care, a way to evaluate if standards of care are being followed, and if service is efficient [23]. Additionally, care pathways should be dynamic, constantly changing with approvals of new treatments, patient preference, a patient's evolving health state, clinician feedback, and localised resource availability and needs.

The definition of a clinical pathway is contentious, with many interchangeably using the terms clinical pathway, patient pathway, and even care plan [19]. The System of concepts to support continuity of care (ContSys) international standard model provides a comprehensive list of terminology related to pathways and continuity of care, including the relevant data points needed to record the concepts in a patient care record [24]. ContSys can be used to help standardise terminology around care pathways and care plans between research disciplines.

There are eight terms of particular relevance to this thesis:

- **Indirect event** or 'indirect healthcare activity period' involves one or more healthcare professionals completing a healthcare activity without the patient being present. An indirect event could be an MDT meeting to decide upon a treatment plan for a patient or the preparation of a prescription. In a care

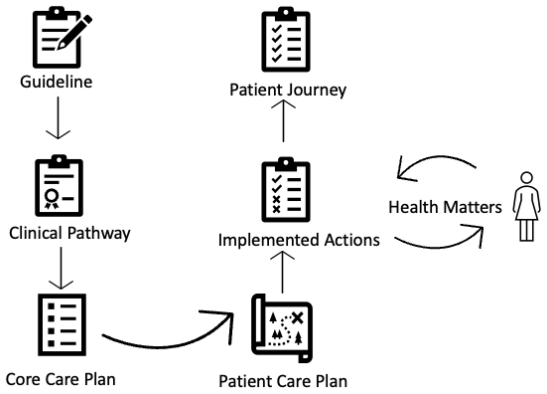


Figure 2.1: A visualisation of the ContSys terminology and how pathways, plans, and actions fit together.

record, they provide information regarding the administrative delivery of care pathways and plans.

- **Direct event** represents an interaction between a patient and a healthcare professional. This includes appointments, surgical procedures, and wound dressings.

- **Healthcare activities** are activities intended to directly or indirectly to improve or maintain a health state. They can be indirect or direct events.

- **Clinical guideline** a set of statements that assist healthcare professionals in deciding which healthcare activities to perform for a specific health issue; these are generic and do not concern an actual patient.

- **Clinical pathway** a workflow for care related to a specific diagnosis, clinical trigger, or symptom; a general plan, typically set out at a national level, that reflects best practice based on clinical guidelines and is applicable to all patients. It corresponds to the set of all allowed journeys.

- **Core care plan** a refinement of the clinical pathway narrowing the options to reflect the approach taken by a particular organisation; sets out treatment intentions for a group of patients. The corresponding journeys will be a subset of those allowed for by the clinical pathway.

- **Patient care plan** a further refinement, tailored to the needs of an individual patient, reflecting health state, comorbidities, and treatment preferences, and describing future care intentions; may be updated at any time. The corresponding journeys should be a subset of those from the core care plan.

- **Patient journey** a sequence of healthcare activities for an individual patient consisting of direct events drawn from the care record. The corresponding journey should be one of those allowed for in the patient care plan, as a journey should consist of the implemented actions from the care plan. Patient journeys are the actual sequence of events that patients experience during diagnosis, treatment, and management of their symptoms and diseases.

Implementation of pathways and analysis of them can improve patient care by helping to understand and answer questions such as:

- What is the most frequent journey for a patient with a particular disease or symptom?

- What are the most and least frequent variations of that journey?

- Are those variations due to clinical differences between patients, differences in localised core care plans, or is there inequitable delivery of care?

- To what extent are these patient journeys meeting prescribed care plans and related clinical pathways?

- Are there differences in outcomes between differing journeys? Can this inform future guidelines?

These questions, however, can only be answered if there is an accurate and interpretable recording of clinical guidelines, clinical pathways, care plans, and patient journeys. Electronic health data and patient care records are a rich source for mining care plans and patient journeys [15], [25].

2.3 The Patient Care Record

A care record is a record maintained by healthcare organisations for the purpose of providing care to an individual patient. This is distinct from a personal health record maintained by a patient or carer [26]. The data a care record contains may be messy and fragmented, created at different times by different people using different systems for purposes unrelated to research or systems analysis. As Goldacre and Morley [27] observes, it is "historically and by design" an *aide memoire*: a practical record to help manage care.

A care record will contain a wide range of documents: notes, forms, reports, letters, observations, results, prescriptions, discharge summaries and more. When electronic, these are stored in electronic patient record (EPR) systems; their output is referred to as electronic health records.

The primary purpose of a care record is to support a patient's continuity of care as they move across the system by providing care professionals with a history of observations and recommendations, clinical decisions taken, and interventions performed. The information recorded is intended for manual interpretation, one patient at a time, by care professionals working in a similar context familiar with the documentation conventions and practices of the organisations involved. They may also have access to the patient, who may be able to clear up any ambiguities in the record.

An additional purpose of a care record is for reimbursement. For reimbursement, the care record is translated into a coded form, with each diagnosis, medicine, or procedure mapped to a standard term of a specified medical terminology in a process called clinical coding [28]. Each code will have a financial amount attached, allowing for efficient remuneration. The transformation of care record into clinical codes for reimbursement in the NHS is not generated or reviewed by clinicians but instead produced by administrative staff called clinical coders [29], [30]. This data is collated and available for secondary research in the form of the Hospital Episode Statistics [31], further described in Section 2.8.

2.3.1 Clinical Coding

Clinical codes are typically alpha-numeric identifiers representing a particular concept in a coding system. In theory, coding systems standardise clinical records reporting into terms that can be statistically analysed, enabling comparison between hospitals and countries. In practice, healthcare systems encode their data in a wide range of classifications, terminologies, and ontologies using personalised rules that vary between, and even within, countries and healthcare systems [32].

A terminology is the systematised nomenclature of a specific community of practice, as determined by a centralised body or common usage, which ensures consistency and clarity of terms [32]. Clinical terminologies are systematised nomenclature for defining diagnoses, symptoms, complaints, medications, and clinical procedures.

An ontology is a terminology that uses formal semantic rules to define relationships between and within concepts. Semantics define the meaning of words, phrases, or concepts. These semantic relationships create a type of descriptive logic that can then be reasoned or searched over [32]. More simply, an ontology is a classification system that defines the properties and concepts of a particular subject area and how these concepts are related to each other using explicit expressions.

Ontologies differ from statistical classification terminologies in that they define relationships between concepts in a way that allows logical reasoning, enabling the inference of relationships not strictly defined in the terminology [32] . For example, if the term “myocardial infarction” is semantically labelled with “is ischemic heart disease,” then a diagnosis of “myocardial infarction” implies the patient also has “ischemic heart disease.”

In England, clinical data is largely reported using three terminologies:

- **ICD (the International Classification of Diseases)** is an international standard maintained by the World Health Organization [33]. It describes diseases, signs, and symptoms and is used worldwide to enable comparison of mortality rates and causes. ICD codes can be between 3 and 7 characters

long. The first three characters, a letter followed by two numbers, indicate the chapter and broad pathology of interest. The 4th - 7th number add more detail. For example, ICD-10 code C18 stands for *malignant neoplasm of colon*, while code C18.4 adds the context of *malignant neoplasm of colon: transverse colon*. In NHS England, diagnosis codes are typically recorded using ICD [34], [35]. It is a statistical classification as there is no other defined relationship between concepts except between parent and child, and the classification is exhaustive as it includes residual categories such as "unspecified" and "other" [32].

- **OPCS** (originally the Office of Population Censuses and Surveys' Classification of Interventions and Procedures) is the procedure coding system created and used by the NHS [35], [36]. OPCS codes consist of one alphabetical character denoting one of 23 "chapters" organised by anatomical site, followed by two numeric digits representing a subcategory, and an optional third digit for further, more specific sub-types. For example, H13 is *bypass of colon* while H13.4 is *bypass of colon by anastomosis of transverse colon to rectum*. OPCS, like ICD, is a statistical classification [32].

- **SNOMED CT (Systematised Nomenclature of Medicine Clinical Terms)** is an international standard maintained by SNOMED International containing over 350,000 medical concepts covering diseases, symptoms, procedures, medications, body structures, and even social determinants of health [37]. Each concept is described by polyhierarchies and semantic relationships, making SNOMED an ontology. Polyheiarchy means each concept can have one or more parent or child concepts. Semantics describe the relationship between concepts. Examples of these semantic relationships include "is a", "associated with", "caused by", and "finding site". SNOMED's concepts and their relations can be reasoned over using logical rules described as queries. The SNOMED standard includes a language for this exact purpose: the Expression Constraint

Language (ECL). In NHS England, all GPs are required to use SNOMED, and many hospitals are also adopting it.

It is important to note that these codes will not generally form part of the care record used by clinicians. Clinicians do not make decisions based on codes, but on the amalgamation of information provided by the patient directly and the results of follow-up testing. Therefore, health data created using clinical codes typically lacks the context to understand why a procedure was conducted or a diagnosis was given. At the same time, the original record contains other, more detailed information about the care delivered that could be useful in healthcare research [30]. Using electronic health data, including patient care records directly, for research purposes requires knowledge of the data's provenance, data specification, and an understanding of how the data is stored in the EPRs.

2.4 Learning Health Systems

A learning health system is a healthcare system capable of learning from every patient experience to deliver continuous evaluation, improvement, and innovation [5], [38]. Learning health systems are formed through a virtuous cycle of turning practice into data, data into knowledge, and knowledge into practice. The technical building blocks that support this cycle include effective and automated data capture, a focus on data quality and provenance recording, data storage and access, information governance, and interoperability.

The rationale for developing learning health systems can be varied, but many focus on improving patient experience and outcomes, reducing unwarranted variation in patient care, and optimising the application of knowledge [38], [39].

Provided that EPRs hold the information needed for adequate characterisation of patient journeys, learning health systems can address ineffective delivery of care and improve outcomes [39], [40]. In the context of a learning health system, a care pathway represents the knowledge, obtained in part from data, that is used to inform practice. By examining EPR data, it should be possible to determine

whether a patient's journey is following a prescribed care pathway; if it is not, then there may be opportunities to improve the quality and effectiveness of care by enabling closer adherence to recommended practice, or the pathway itself may require updating to take account an advance in our understanding of disease, a change in policy, or an oversight in the original pathway design. One method to support the transformation of data into knowledge is process mining.

2.5 Process Mining

Process mining is a set of techniques for converting event data into process diagrams to represent what is happening in the real world, and then compare those process diagrams with idealised models or process descriptions to determine areas of variation [9], [10]. It is a bridge between data mining and business process modelling and analytics [11]. In the process mining literature, it is stated that there are three main aims or types of process mining: process discovery, conformance checking, and process enhancement [11], [41]. They are represented in Figure 2.2.

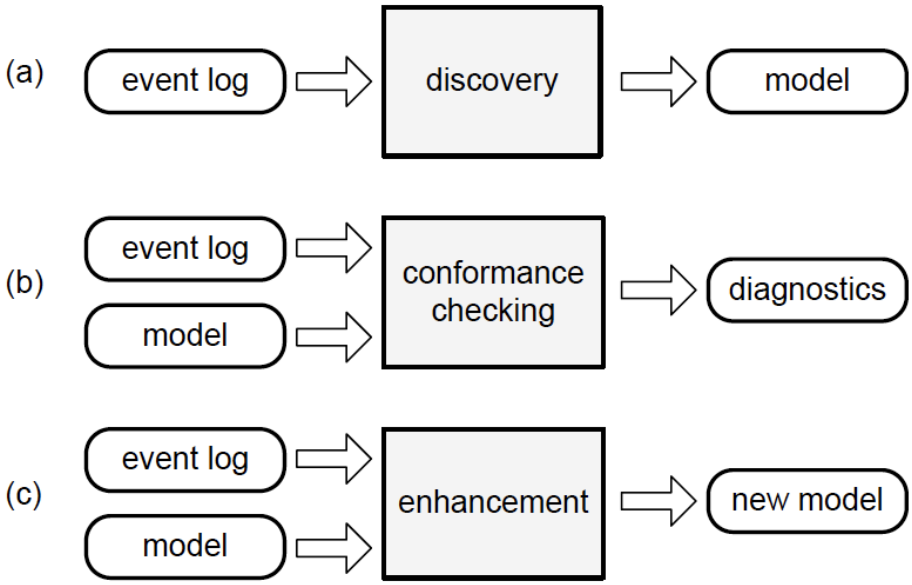


Figure 2.2: Figure taken from van der Aalst, Adriansyah, de Medeiros, *et al.* [11] visually explaining the three main types of process mining, (a) process discovery, (b) conformance checking, and (c) process enhancement, in terms of input and output. The model output from discovery is the same model input for conformance checking and enhancement.

Process mining is applicable to a wide range of systems and sectors. The only requirement is that the system has a mechanism for recording the behaviour of its processes. Extraction and transformation of the data generated by the process into an analysable format is called event-log generation. Event-log generation, akin to data cleaning in data science methodology, is typically the longest step in process mining.

2.5.1 Event-log Generation

Event-logs are ordered sequences of events with an associated timestamp, activity name, and case identifier. It is important that each event refers to a well-defined step, or activity, in the process (e.g. a blood test) and is related to a particular case (e.g. a patient). Event and activities are terms often used interchangeably in process mining, and in this thesis an event is an implemented clinical activity.

Event-logs should reflect the general order of a process and the most frequent pathways taken [41]. A specific sequence of events for a particular case is referred to as a trace; therefore an event-log is a multi-set of traces [9]. Additional information such as the performer of the event (i.e. the doctor performing the test), the timestamp of the event, or additional data elements related to the event or case (e.g. cost of the surgery, age of the patient) may also be stored in event-logs [9], [11], [42].

The case notation of available data will determine the events included in an event-log and what processes can be followed. Possible case notations in the healthcare context include patient ID, where all events attached to a patient will be included, and admission ID, where all events associated to a specific hospital admission will be included. The traces in an admission case-based event-log will be a subset of a patient ID case-based event-log.

Event data attributes in a healthcare context include the results of a blood test, with the withdrawal of blood for the test being the event, or the weight of a patient, with the event being weighing the patient or performing a wellness check. Event-logs are the basis for process mining to discover, monitor, and improve real-world processes [10].

2.5.2 Process Discovery

Process discovery involves applying algorithms to event-logs to construct a model of what is happening in the system without any previous information [9], [42].

When generating and critiquing a discovered process model, there are four things to consider: fitness, simplicity, precision, and generalisability [41]. This means that a discovered process model should be able to explain an observed behaviour (fitness) as simply as possible (simplicity), while not presenting any behaviours that are unrelated to the event data (precision), while still being generalisable to all situations present in the event-log (generalisable). Discovered models "should be treated as purposeful abstractions of reality". This means that different models should be created for different stakeholders to ensure they solve the users' specific needs [11].

There are four perspectives that a process model can take: control-flow, organisational, case, and time [9], [11]. The control-flow perspective focuses on the ordering, or flow, of cases between activities, with the goal to define a model that characterises all possible paths. The organisational perspective focuses on how resources responsible for executing actions are related. The case perspective defines models based on the properties of cases, which can include related data elements or on the case's specific journey within a process. The time perspective generates models focused on the timing and frequency of activities in a process [41].

Methods for process discovery include process mining discovery algorithms, such as the heuristic miner, inductive miner, and the fuzzy miner [41]. Discovered processes can then be expressed using different notations such as petri nets, business process model and notation (BPMN) diagrams, and directly-follows graphs (DFGs), as shown in Figure 2.3 [43].

These algorithms can automatically derive process models which can be refined manually with domain knowledge to improve their representations [41]. New techniques are being developed to handle specific process discovery issues, such as local process model discovery to break-down large-complex processes, and clustering techniques to discover similar trace variants within a larger process [44].

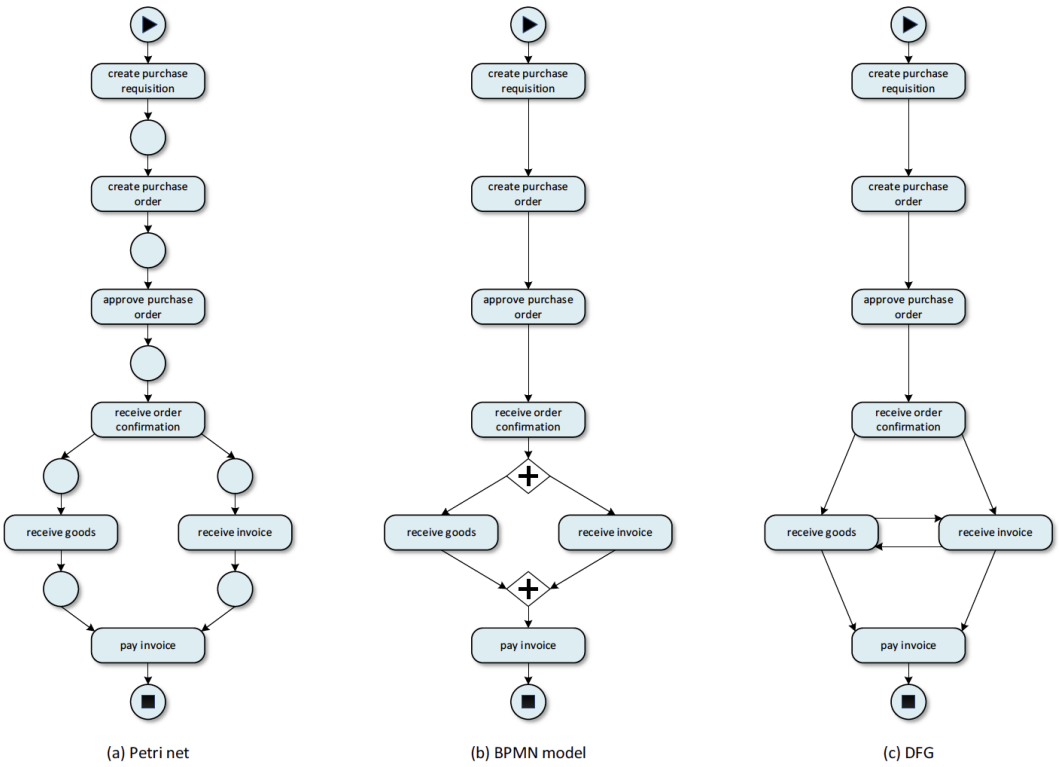


Figure 2.3: Figure taken from van der Aalst [43] showing the differences between three process model representations (a) petri nets, (b) business process model and notation (BPMN) diagrams, and (c) directly-follows graphs (DFGs). The processes discovered are for a Purchase-to-Pay (P2P) process.

In this thesis, discovered processes are represented using directly-follows graphs, for example in Figures 3.3. Directly-follows graphs visualise the flow of cases (e.g. patients) between activities. The circles, or nodes, represent each activity present in the event-log. The value inside the node is the percentage of cases which complete that activity at least once. The arrows, or edges, show how the percentage of cases that move from that activity to the following activity. Certain orders can be inferred from the arrows, including starting activities and typical final activities. The DFGS in this thesis are generated directly from the data, without using a process discovery algorithm first, using the bupaR package in RStudio [45].

2.5.3 Conformance Checking

Conformance checking compares processes to determine areas of variation. Examples include comparing a discovered process against its idealised model to determine if

the process is being delivered as intended, or comparing discovered processes from different organisations to determine variation in process delivery [9], [11].

There is no standard technique adopted for conformance checking; selection of method is based on the goal of conformance checking, the availability of event-logs, and the representation of the process model [41]. The most commonly used techniques include "replaying" an event-log's data on a discovered or idealised process to identify deviations, aligning traces to the discovered model, and manual checking and filtering of specific traces against specific performance metrics.

2.5.4 Process Enhancement

Process enhancement involves the annotation of discovered models with additional information such as activity duration, waiting time information, organisational and resource information, and decision logic. The enhanced models can be used to identify areas of potential improvement. For example, identifying deviations in the discovered processes such as bottlenecks, fraud, or even identifying process variants that are more efficient than the idealised processes. The impact of any resulting changes to the system can then be evaluated through further process mining analysis [9].

2.6 Process Mining for Healthcare (PM4H)

Process mining techniques have been applied to the discovery of patient journeys and to the determination of the extent to which these journeys conform to intended care pathways [13], [19], [44], [46], [47].

Although the core algorithms and techniques of process mining are domain agnostic, their adoption within healthcare remains relatively limited in comparison to other sectors [15], [41], [48]. The main explanation for this is the heterogeneity of healthcare data and of the healthcare processes themselves [47].

Healthcare data may be created and stored in a range of different settings using different systems and data entry practices, and then transformed in different ways prior to analysis including into clinical codes. This data is not recorded

in a process centric way and therefore extensive pre-processing of the data into analysable event-logs is required. This is time-confusing and often not straight forward, especially as many publicly available healthcare datasets, do not explain how the data is created [49]. Additionally, there are challenges presented by the ways in which the data is represented in EPRs. The forms, tables, columns, and fields used to hold the same kind of information may be different in different systems, and in different configurations of the same system. This increases the amount of resources required to extract useful information for research and even basic hospital management tasks [50], [51].

Furthermore, patient journeys and healthcare processes are human-centric, driven by preferences and ongoing discussions between patient and clinician that may not be fully recorded in health records, as well as changes in treatments and clinical knowledge. This often results in the discovery of "spaghetti processes": process models that are too dense and complex to be easily visualised and interpreted [9]. All of these complexities of the healthcare sector require personalised techniques within process mining [42], [52].

Another factor is the lack of a standard modelling language for care pathways, which prevents efficient conformance checking. While notations such as Business Process Modelling Notation (BPMN) have been used in the literature, existing practice often relies upon a mixture of free text and bespoke 'flow chart' notations: for example MD Anderson Cancer Center's clinical algorithms [53], [54]. These flow charts require individual interpretation to create an idealised, computable model for conformance checking. There is a shift in the sector, however, to making clinical guidelines and care pathways computable and machine readable, shown by the World Health Organisation's creation of "Digital Adaptation Kits" and the US CDC's "Adapting Clinical Guidelines for the Digital Age" initiative [55], [56].

These challenges are highlighted by Erdogan and Tarhan [41], who conducted one of the field's most extensive systematic reviews, encompassing 172 papers on the quantity, type, and results of process mining research in healthcare. They identified a gap in process modelling of clinical pathways (as opposed to administrative

pathways) and of processes involving multiple departments or hospitals; 61% of the reviewed papers modelled an organisational process in a single department of a single hospital. All processes in this thesis analyse patient journeys across multiple departments. Chapter 3 focuses on journeys within in a single hospital while Chapter 4 models journeys across multiple hospitals.

2.7 Process Mining and Clinical Research

The main challenges of process mining in healthcare – unstructured data, spaghetti-models, and lack of translation to the clinical setting – can be addressed through closer alignment of traditional clinical research methodologies and process mining.

Epidemiology, a branch of clinical research, is focused on measuring the distribution and determinants—the "who, where, when"—of disease and other health-related events [57]. At its most basic, it describes the rates of occurrence of a disease, noting differences in groups based on who they are, where they live, and when they lived. It can also determine risk factors of a disease and measure their effect on health outcomes such as life expectancy. The focus is on counts and snapshots, rather than flows and journeys which more accurately capture the true experiences of patients.

Epidemiology and clinical research methods this thesis considers include funnel plots for comparing institutional performance [58], Kaplan-Meier curves for comparing relative survival risk [59], and logistic regression for understanding the impact of various variables on receiving treatment [17].

Through use of these methods, including having a greater understanding of previous research findings, processes of interest can be determined, and the relevancy of process mining results enhanced. No previous process mining for health literature has explicitly explored the possibility of integrating these two fields to expand healthcare process research.

2.8 Colorectal Cancer Care in the English NHS

Colorectal cancer is cancer of the large bowel, which includes the colon and the rectum. The colon is the part of the digestive system that absorbs water and removes food waste from the body. The rectum is the part of the body which stores stool until it is ready to be passed. It is the fourth most commonly diagnosed cancer and the second most common cause of cancer death in the UK [60]. Symptoms of colorectal cancer include a persistent change in bowel habits, blood in stool, abdominal pain, and unexplained weight loss.

NHS England has recently changed many of its cancer pathway waiting time standards. This includes implementing a new diagnosis timed route that moves from the 2-week wait, which stated that patients should be seen by a specialist within two-weeks of an urgent referral, to the 28-day faster diagnosis standard, which states that a patient should receive a diagnosis within 28 days of an urgent referral [61]. Additionally, there are time targets for starting treatment. These include no more than 62 days from urgent referral for suspected cancer and no more than 31 days from the date of treatment decision [62]. These changes create a need to monitor the effectiveness of these pathways in real-time and ensure quick uptake and adherence to new guidelines.

In October 2024, only 68% of NHS cancer patients were treated within 62 days of an urgent referral– significantly lower than the national standard of 85% of patients to be treated within 62 days. However, 77% of patients were told they do or do not have cancer within the 28-day target, which is above the target of 75% of patients. 92% of patients received their first cancer treatment within one month of a decision to treat, slightly below the target of 96% [63].

Waiting times for colorectal cancer are lower than the national averages for all cancers: only 62% of patients are diagnosed with 28-days; 90% start treatment within 31 days of a decision to treat, but only 56% start treatment within 62 days of urgent referral [64].

The National Institute of Health and Care Excellence (NICE) sets clinical guidelines, including for colorectal cancer treatment, at a national level for England

[65]. NHS England uses these guidelines to create clinical pathways, but each NHS Trust can modify these guidelines and pathways for their local area, and even professional bodies can issue their own recommendations for clinical pathways.

Treatment of colorectal cancer is dependent on the location (colon or rectum), stage, and size of the cancer, if/where it has spread in the body, as well as patient preference and physical health. NHS guidelines state that everyone diagnosed with colorectal cancer should be under the care of a multidisciplinary team [66]. A multidisciplinary team is a team of different health professionals who meet weekly to discuss cases and decided the best course of treatment for a patient. There are 120 NHS colorectal cancer MDTs in England.

MDTs, NHS trusts, and other health and social care organisations are further organised into Cancer Alliances. Resources within a Cancer Alliance should be working together to deliver equitable and effective care [67]. There are 21 Cancer Alliances in NHS England.

2.9 English Cancer Care Datasets

The incidence of cancer in its various forms, the contribution that it makes to overall mortality, and the cost of treatment have made it the subject of national disease registration programmes aimed at monitoring and improving care delivery and outcomes. This thesis uses multiple data-sources that are routinely used in research to study cancer care delivery in England.

National Cancer Registry

The National Cancer Registration and Analysis Service (NCRAS), which is a part of the National Disease Registration Service, defines the gold-standard "date of incidence" for all patients diagnosed with cancer in England. It does this by collecting submitted data from hospitals and MDTs across the country and then curating the data into relevant data-points and datasets [68]. These additional datasets include the Systemic Anti-Cancer Therapy data set (SACT) and the

Radiotherapy Data Set (RTDS), which can be linked to other NHS datasources for longitudinal follow-up of patients.

The Systemic Anti-Cancer Therapy (SACT) dataset contains detailed information on the prescription and administration of systemic anti-cancer therapies, such as chemotherapy and immunotherapy, across all NHS England trusts [69]. The Radiotherapy Dataset (RTDS) contains detailed information on the planning and administration of radiation therapy across all radiation care providers [70].

Hospital Episode Statistics

In England, each hospital submits monthly to the NHS detailed, OPCS coded information on inpatient, outpatient, and emergency care delivered for the purpose of reimbursement. This is made available for secondary purposes in the form of the Hospital Episode Statistics (HES) dataset [71].

The HES dataset is subject to the limitations mentioned in Section 2.3: there is little in the way of context and only reimbursable codes are reported even if other activities occur. However, it does add significant value to the registry datasets: not only can surgical procedures for cancer treatments be evaluated, but the data can also reveal whether and how patients are being treated for other concurrent conditions.

CORECT-R

For this thesis, the NCRAS, SACT, HES, and RTDS datasets were accessed using the Cancer Research UK funded COloRECTal cancer data Repository (CORECT-R) [72]. It is a secure data environment (SDE) bringing together multiple routine datasets relevant to colorectal cancer in the UK to be made available to requesting researchers.

A coding catalogue is maintained and published for CORECT-R, explaining how clinical codes are aggregated to define primary treatments such as major resection and comorbidities such as diabetes. This catalogue is essential to the validation and reproducibility of analysis as it provides an understanding of the data transformation from source. It also provides a basis for suitable activity abstraction needed to generate clinically relevant event-logs from the data.

NHS Trust EPRs and EHRs

This thesis additionally uses data sourced directly from a regional NHS trust's EPR systems. This data is accessed using the Thames Valley and Surrey's secure data environment [73], which collates data from across the Trust's EPR systems to provide a picture of care delivery. Data provided to researchers is based on a specific data schema request for a specific research purpose. The specific cohorts and EPR systems utilised in this thesis are discussed in the corresponding chapters.

"In the longer run and for wide-reaching issues, more creative solutions tend to come from imaginative interdisciplinary collaboration." - Robert J. Shiller

3

An Epidemiology-informed Process Mining Approach

Contents

3.1 Introduction 25
3.2 The Approach 28
3.2.1 Defining a Research Question 29
3.2.2 Literature Review 30
3.2.3 Patient Population and Subcohorts 31
3.2.4 Event-log Generation 32
3.2.5 Mining for Processes and Comparison of Results 33
3.2.6 Review Results and Iterate with Domain Experts 33
3.3 Colon Cancer Treatment Journeys at a NHS Trust . 34
3.3.1 Research Question and Literature Review 34
3.3.2 Patient Population and Subcohorts 35
3.3.3 Event Log Generation 37
3.3.4 Results 37
3.4 Importance of Data Provenance 49
3.5 Discussion 50

3.1 Introduction

Various reviews and publications have been written about the challenges of process mining in healthcare and have provided future work recommendations for the PM4H community [10], [13], [15], [42], [47], [74], [75]. Many of their challenges and

recommendations overlaps, and a summary of the ones this thesis considers are:

- Establish a standardised terminology to support process mining in healthcare which provides clear descriptions of process mining concepts within the healthcare context, and where possible, explicitly define the relationship to existing terminologies in the healthcare domain.
- Start from real-world problems by working with clinicians and staying up-to-date with the latest healthcare research.
- Consider healthcare specificities during technique development. These include considering that the flow of a clinical process is heavily influenced by patient conditions not directly related to the process or that infrequent behaviours should be considered in process discovery.
- Provide holistic, multi-disciplinary process views by incorporating all sources of process-related data which are available, even if in an unstructured format.
- Develop a comprehensive methodology that takes into account the specificities of healthcare and clinical processes, including the identification of process-related questions, preparation of process execution data, and a suitable process mining technique to answer the question and address challenges posed by healthcare data.
- Look at the process through the patient's eyes to model the patient journey is currently under-researched in PM4H, which often focuses on the organisational perspective.
- Domain experts should be involved at every stage of process modelling to ensure that research is translated into action, including in defining the research goals and informing activity selection.
- Establish technologies to incorporate the process perspective into healthcare information systems such as EPRs.

- Develop multi-perspective approaches to combine the control-flow with a resource and/or time perspective and consider relevant key performance indicators that can be measured using these multiple perspectives.

To address these challenges and recommendations I present the epidemiology-informed process mining approach. The goal of epidemiology-informed process mining is to mine patient journeys from EHRs to determine if there is variation in care between patient groups. It does this by providing a flexible, generalisable, and reproducible approach that is aligned with clinical research terminology to ensure clinical relevancy of insights; the focus is to add context about the patient, clinical guidelines, and their care records into process mining directly. Additionally, the method specifies how to iteratively and flexibly determine which process variants or patient cohorts to investigate; a gap not addressed previously.

The epidemiology-informed process mining approach extends Cremerius, Pufahl, Klessascheck, *et al.* [76] with additional steps from epidemiology and clinical research methodology [57], [77] to create an end-to-end process mining method tailored for the healthcare sector.

Cremerius, Pufahl, Klessascheck, *et al.* [76] describes a reproducible framework for generating event-logs in the widely-used, publicly available dataset MIMIC. The framework includes an event-log generation method, an event hierarchy, and a Python tool for event-log extraction. The method has six steps which starts with stating a goal, followed by defining patient cohorts and case notation, then selection of case attributes, event types and their attributes occurs, finally concluding with enriching event attributes.

Epidemiology-informed process mining also takes inspiration from Cremerius, Patzlaff, Rahn, *et al.* [78], which presents "Data Based Process Variant Analysis", an approach which "allows comparing process variants based on differences in event data attributes by building on statistical tests". The approach shows how to consider differences in patient cohorts and event data attributes during process variant analysis. It also demonstrates an approach for incorporating additional contextual information from the care record, focusing specifically on human measurements such

as blood test results and blood pressure. However, many healthcare datasets don't have these attributes available to them, including the Hospital Episode Statistics [79]. Even in situations where the data is available, linking these attributes to clinical events can be time-consuming and difficult.

Section 3.2 describes each step of the approach in detail. Section 3.3 shows how the approach can be applied to the analysis of treatment processes by investigating differences in colon cancer care delivered by a major English regional NHS Trust to provide insights to clinicians and administrators on if they are delivering equitable care.

The approach and case-study in this chapter was originally presented at the 2024 Process Mining in Healthcare mini-track at the Hawaii International Conference on System Sciences, with the corresponding paper published in the conference proceedings [51]. I have developed and expanded the approach as well as re-completed the analysis for this thesis, with the figures produced by my co-author Owen Dwyer.

3.2 The Approach

The steps of the epidemiology-informed process mining approach are:

1. Define the research question and process for investigation
2. Conduct a literature review
3. Define the patient population and subcohorts
4. Identify the events of interest
5. Generate event logs and mine for processes
6. Compare results to what is expected
7. Investigate deviations
8. Present findings to domain experts and iterate through the approach

3.2.1 Defining a Research Question

In this approach, research questions should concern a particular patient cohort with a particular health issue, as this ensures closer alignment to clinical guidelines and care plans, which are developed on the same basis. This additionally facilitates conformance checking and leads to process models that clinicians can interpret qualitatively.

The research question should be driven by a clinical need or a gap in the research literature. These needs and gaps can be suggested through discussion with a domain expert or discovered during the literature review. Any epidemiology-informed process mining results should be tailored to the target audience to ensure interpretability and relevancy. For example, a hospital administrator may prefer to see a high-abstraction process focusing on time-based performance. In contrast, a clinician may prefer to see a low-abstraction process focusing on control-flow performance.

Another consideration for the choice of research question is the nature of the available data. For example, in the widely-used MIMIC dataset (a publicly available database from the EHRs of a major USA based hospital), all data associated with a particular hospital admission is grouped under a single identifier [49]. It also contains data on vital signs, such as blood pressure and medications dispensed. The English Hospital Episode Statistics (HES) dataset, on the other hand, has a separate identifier for each consultant episode of care within a different identifier for the complete hospital admission; each patient also has a unique, pseudonymized identifier [68]. It only contains information on coded surgical appointments and does not include the value of vital signs taken or the name of any medications dispensed; for example, only the administration route of a generic drug band such as“chemotherapy” is recorded. Such differences between these datasets can affect not only the choice of case identifier for event-log generation, but the scope of the research question as well.

An example of effective research question driven healthcare process mining is Pijnenborg, Verhoeven, Firat, *et al.* [44]. They centred their research question on the

issue of palliative care treatments for oesophageal and stomach cancer patients. This allowed them to determine common practices in palliative care, link them to survival outcomes, and provide evidence to inform the development of clinical guidelines.

3.2.2 Literature Review

In epidemiology, literature reviews are an important part of the process and are often required in published research papers. A literature review should provide a comprehensive over-view of what other research concludes about the question of interest. For example, providing answers to what are common treatments for a disease, factors that impact disease outcomes or treatment choice, and what are the national and international guidelines and pathways for disease diagnosis or treatment [77].

The literature review will provide a basis for the assessment of findings, indicate contextual information that may be relevant to include in process discovery, and suggest how to define population cohorts and subcohorts. These reviews are often conducted using research databases such as PubMed and MedLine.

Additionally, having a literature review will provide further evidence base for the relevancy of the process mining results, and will bolster credibility within the clinical research community. It will help contextualise the importance of the results and highlight how process mining can address gaps in the clinical research literature.

Taylor, Swinson, Seligmann, *et al.* [80]'s paper addresses an identified gap in the colorectal cancer treatment literature– the limited evidence of variation in the patterns of chemotherapy usage for English colorectal cancer patients. Their research filled this gap by quantifying variation in post-operative chemotherapy use for patients of a specific cohort. The results of the paper however highlight a necessary gap in epidemiology that process mining can fill - quantifying the usage of treatments across a whole pathway instead of just specific treatments provided at specific moments in the pathway.

3.2.3 Patient Population and Subcohorts

Existing approaches to reduce process model complexity [75] are focussed upon event-logs and process traces: filtering or grouping events, or clustering traces to form trace variant groups for separate analysis [15], [74], [81]. Contextual information from the care record and information about care guidelines discovered during the literature review provide another means of reducing complexity by stratifying the dataset into cohorts of patients with similar characteristics who are more likely to have similar journeys.

Suitable data points for stratification can be determined during the literature review. For example, factors that have previously been identified to affect the clinical pathway of interest, or factors that have not be researched yet and are identified as gaps in the literature. Common factors that impact care pathways include age, sex, socio-economic background, and patient comorbidity [57]. Some of these data points may be directly recorded in a patient's care record, while others may need to inferred from the care record indirectly. For example, identification of comorbidity or surgical complication can happen by identifying the occurrence of events outside the pathway of interest.

When presenting results process miners should present the literature that informed their patient population definition, selection of events, and other factors involved in generating the process insights. They should also present the demographics and population statics of the researched patient population as is typically done in clinical research, because it provides additional context for interpretation of results. Patient population statistics can include how many patients are in each discovered process, at what age, in which sex group, and any comorbidities that are documented or other demographic factors that are known such as ethnicity and social deprivation.

Baker, Dunwoodie, Jones, *et al.* [46] provide an example of cohort definition for process mining, based upon cancer type, chemotherapy drug, and the intent of the chemotherapy treatment. Their work also involves the inclusion of health states—additional contextual information—as events in the model: for example, whether

a patient's white blood cell count was too low for chemotherapy. The resulting discovery phase of patient journeys during chemotherapy is clear and concise.

3.2.4 Event-log Generation

Selection for specific activities and events of interest is essential when analysing patient journeys. This is because patients can be on multiple, concurrent pathways– for example being diagnosed with cataracts while being treated for colorectal cancer–resulting in events that happen at similar times but for completely unrelated processes. This differs from traditional business process mining which often has a restricted number of activities that correspond to the same process. Determining which activities are relevant to the care pathway and process of interest, therefore, is paramount for ensuring relevant and focused analysis.

As a starting point, clinical guidelines and care plans provide a core list of activities that should occur in a patient's journey. These lists may prove sufficient for population-level analysis focusing on the most and least frequent journeys, the average timing between events, and the number of activities involved. Individual-level analysis may require additional events representing comorbidities or health states preceding or following the initial event-log.

Another method for activity selection can be to look at previously published clinical code-lists for the disease of interest. These code-lists will provide starting points for identifying relevant events from patient care records. The UK Colorectal Cancer Intelligence Hub's COloRECTal cancer data Repository (CORECT-R) project published a publicly available code-list mapping NHS OPCS clinical codes to colorectal cancer surgery procedures such as major resections and stoma formation [72]. This list can be used by anyone who has data coded in OPCS terminology and wishes to investigate colorectal cancer treatments for their own research purposes.

Once the relevant activities have been determined, a data dictionary containing the codes and their mapping to corresponding activity categories should be created and then published with any findings. This will make healthcare process mining results reproducible and comparable across analyses.

Once a patient cohort is defined with a corresponding activity list, the activities of interest can be searched for in the data, and an event-log can be generated. The event-logs can also be analysed at different abstraction levels. Abstraction is the act of grouping different activities into larger categories: for example blood test and x-rays being grouped into a higher-level activity named diagnostic tests. Analysing event-logs at different abstraction levels will provide different insights into care processes and their deviations. These various abstraction level mappings should also be included in any data dictionaries created and published.

Chapter 4 will introduce additional healthcare process mining methods for patient pathway activity selection and event-log generation and will present a new activity selection method which leverages the knowledge embedded in medical ontologies.

3.2.5 Mining for Processes and Comparison of Results

Once the event-log is generated a process model can be discovered using the variety of techniques available in process mining.To gain an adequate understanding of a particular issue, both population and individual level analysis, alongside low and high granularity of events, may be required. Directly-follows graphs may be enough to answer the questions asked, but additional analysis such as dotted charts, logistic regression, and Chi-squared tests may also be useful.

For the application of this approach in Section 3.3, only directly-follows graphs are used. This is due to their easy comprehension; they don't require knowledge of process discovery algorithms, and they avoid many of the problems of process discovery algorithms including complete reduction of the model due to high variability of patient pathways [9].

3.2.6 Review Results and Iterate with Domain Experts

Once areas of interest and deviation from expected results, based on the literature review, are discovered, they can be discussed with domain experts. Multiple perspectives (e.g. time, resource, patient flow) along with additional context may be needed to understand the significance of deviations from a patient care plan, a

core care plan, or a clinical pathway. Working with domain experts and the insights' end-users will determine which perspectives are required.

An additional area that domain experts can help with is the refinement of discovered process maps. For example, to reduce loops event-logs can be reduced to only include the first instance of an event. Iteration with domain experts may be needed to increase comprehension of the graphs by removing unnecessary activities, and refinement of the research question may be needed to focus on smaller sections of the pathway.

As new insights arise, new questions will arise in tandem. This iteration and refinement of the research question and the scope of resulting process maps and insights should be done closely with domain experts and end-users to ensure the applicability of results.

3.3 Colon Cancer Treatment Journeys at a NHS Trust

To show how the proposed approach usefully incorporates domain expertise and additional care record data, I investigated colon cancer patient journeys at a major English regional NHS trust.

The dataset used contains anonymous information on the treatment of 1,417 patients with an ICD-10 code of C18 (colon cancer) recorded in the main cancer patient management database at the trust. The data covers a period of 11 years, from 2013 to 2024, and includes all inpatient and outpatient hospital activities coded using the OPCS terminology, whether they are related to colon cancer or not, providing a rich history of events to support the investigation.

3.3.1 Research Question and Literature Review

The research questions were "what are the common treatment journeys for colon cancer patients in the hospital?" and "is there variation in treatment journeys between patient subcohorts?"

During the literature review phase, I reviewed the relevant NICE clinical guidelines for treatment selection, the national clinical pathway defined by NHS England, factors that have been shown to impact life expectancy and treatment choice, and published findings on the variation of treatments for colon cancer across England [17], [62], [65]. Based on the review, I expect the results of this study to show rates of treatment to decrease with increased age, increased Charlson Score, and with increased deprivation [16], [17], [82]–[84].

National guidelines state that treatments for colon cancer include resection to the tumour, systemic anti-cancer therapy (SACT) for high-risk stage II and stage III patients, and occasionally radiotherapy palliatively or for the management of metastatic sites. Treatment selection should be based on tumour stage (information I sadly do not have), patient health state (which can be inferred from the greater care record), and patient preference (another data point I do not have) [65]. NHS cancer pathway guidance states that patients should receive a diagnosis within 28 days of referral for suspected cancer and should start treatment within 62 days of referral [62].

3.3.2 Patient Population and Subcohorts

The patient population is defined as those having colon cancer and treated at the trust. There were two ways to identify patients with colon cancer within the dataset. The first was to identify patients with an ICD-10 C18 (colon cancer) code somewhere in their general hospital out- and in-patient appointment care record, which is the Orbit EPR system for this trust. The second, a more targeted approach, was to identify those with a ICD-10 code of C18 in a care record from a cancer management electronic health record system, such as the MDT meeting record system Infloflex.

When using the first approach, many patients could be identified that do not have confirmed cancer but who had a procedure for suspected cancer, resulting in a patient population of 4,215 patients. The second approach increases the chances that these are patients with confirmed cancer diagnosis, however it reduces the ability to investigate what happens to patients that do not make it to the MDT

meeting; the resulting patient population is 1,571 patients. Figure 3.1 shows a Venn Diagram describing the overlap of these patient groups. For this study, the second approach was utilised to define the patient population as it ensures that patients selected have cancer and are under review by the hospital for treatment.

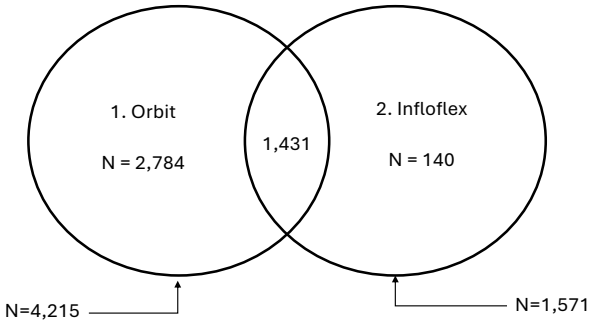


Figure 3.1: A Venn Diagram describing the overlap of the patient groups created by patient cohort definition approach 1 (surgical diagnostic codes from the general hospital EPR, Orbit) and approach 2 (diagnostic codes from the cancer management EPR, Infloflex)

The 2,481 patients identified only through the Orbit EPR system had a diagnosis C18 code attached to an appointment with no procedure, or had a C18 code attached to a diagnostic endoscope procedure or diagnostic imaging procedure (such as CT or MRI). The 140 patients only identified within Infloflex were due to non-C18 ICD codes being used for the relevant diagnostic procedures in Orbit. The ICD codes found were for related sites to the colon– malignant neoplasm of the rectum (C20) or malignant neoplasm of the sigmoid junction (C19)– or were more general for benign neoplasm (D12x) or neoplasm of uncertain or unknown behaviour (D37x).

Patient cohort subgroups were defined based on patient age, Charlson score (a measure of comorbidity), and the Indices of Multiple Deprivation (a relative measure of socio-economic deprivation for a small area surrounding the patient's home address). This was based on the literature review results for factors that impact cancer treatment and life expectancy.

3.3.3 Event Log Generation

The data comes from four different EPR systems within the Trust: Infloflex provides the patient demographics and their cancer date of incidence; Aria Medonc provides information on chemotherapy regimens; Aria Radonc provides information on radiotherapy adminstration; the surgical procedures, major and minor resection, come from the main EPR system Orbit. The data was accessed within a secure data environment and provided for by the hospital.

For this study, only patients with 1-year or more follow-up data from incidence were included, reducing the population by 154 to 1,417 patients. Four treatment types of interest were identified from the care record: major resection, minor resection, chemotherapy, and radiotherapy. The mapping from OPCS codes to this classification of treatments was informed by extensive discussion with colon cancer clinicians and epidemiologists and is aligned with the CORECT-R data coding table, and published to my Github for others to use [85].

For initial event-log generation, only the four treatment types were included as events. An additional event, cancer incidence, was derived by identifying the earliest date for which a C18 code was recorded. Where this incidence date coincides with a treatment date, the incidence is considered to have occurred first. Any duplicated or repeated events (event names in a row) were removed. 60 unique event-log traces were generated for the entire patient population.

3.3.4 Results

Directly follows graphs (DFGs) were generated from the event-log for the entire patient population and for each of the patient cohort subgroups to understand the

flow of patients between activities. A DFG process starts at "start" signifying the nodes directly after are the first in the event-log. The "end" indicates that the node before is the last activity in the event-log. Here, "end" does not signify death, but the end of the patient's record for the events of interest.

If any pathway anomalies were identified in the graphs, the patient's care record outside of the events of interest was re-examined to find context that could form the basis of an explanation. Only DFGs discussed in the results are presented here, additional DFGs for all patient subcohorts can be found in Appendix A.

3.3.4.1 Population Level View

Table 3.1 describes the demographic breakdowns of the colon cancer patient population. As patient care records are written primarily to be read by other clinicians, not all demographic factors are recorded for each patient.

Characteristic	Number of Patients	Proportion (%)
Age Group		
≤49	138	9.7
50-59	149	10.5
60-69	321	22.7
70-79	429	30.3
≥80	380	26.8
Sex		
Female	656	46.0
Male	761	54.0
Charlson Comorbidity Score		
0	938	66.2
1	416	29.4
2	52	3.7
3	11	0.7
Index of Multiple Deprivation (IMD 2015)		
1 - Least deprived	48	3.4
2	90	6.4
3	167	11.8
4	387	27.2
5 - Most deprived	642	45.3
N/A	83	5.9

Table 3.1: Patient characteristics for all 1,417 patients analysed.

Figure 3.2 shows the DFG for the whole cohort of patients. As this is a population-level view, comparison can be drawn with the Cancer Treatments Dashboard created

by National Cancer Registration and Analysis Service (NCRAS) as part of the Cancer Research UK - NHS England Partnership.[86]. The dashboard shows that in England, either individually or in-combination with another treatment, 31% of colon cancer patients receive chemotherapy, 65% have a surgical resection, and 5% have radiotherapy; around 40% of patients receive no treatment at all.

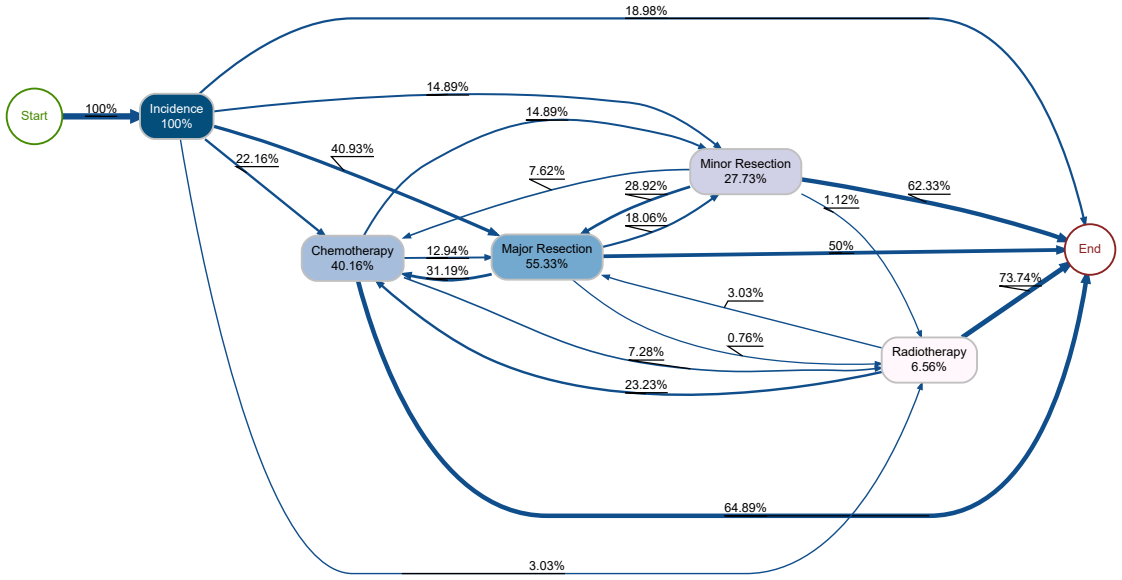


Figure 3.2: Directly follows graph showing treatment journeys for the entire patient cohort of study.

Figure 3.2 shows that the overall distribution of interventions within the hospital reflects, and is even out-performing, these national averages with 40% receiving chemotherapy, 83% having a surgical (major and/or minor) resection, and 7% having radiotherapy; only 19% of patients do not receive any form of treatment. Major resection (41% of patients) is the most common initial treatment, followed by chemotherapy (22%) and minor resection (15%). 41.6% of patients only receive one treatment modality. The 5 most common traces are shown in Table 3.2.

This high-level view of journeys however does not describe whether the distribution of treatments for particular subgroups of patients is as expected. To explore this, the patient cohort is first stratified by age.

Trace	Patient Proportion (%)	Cumulative Proportion (%)
Incidence, Major Resection	20.2	20.2
Incidence (no treatment)	19	39.2
Incidence, Chemotherapy	14.5	53.7
Incidence, Major Resection, Chemotherapy	8.61	62.31
Incidence, Major, Minor Resection	6.70	69.01

Table 3.2: The 5 most common process traces, based on the event-log, for the patient population.

3.3.4.2 Age

Figure 3.3 shows the DFG for patients aged 50 to 59 at the time of incidence, whilst Figure 3.6 shows the DFG for those 80 and over. These show that only 7% of patients aged 50 to 59 received no treatment, whereas the rate for those aged 80 or over is 43%. A large difference in no treatment rates is also seen between those aged between 70 and 79 (Figure 3.5) and those 80 or over, with the no treatment rates being 12% for those aged 70-79.

The biggest difference in singular treatment rates between age groups is with chemotherapy: 71% of patients aged 50 to 59, 53% aged 60-69, and 35% aged 70-79 receive chemotherapy, but only 13% of patients aged 80 or over do.

This can be explained clinically and through patient preference. The range of treatment options narrows with age as the potential benefit, in terms of increased life expectancy, is reduced and the cost, in terms of side effects, is increased due to frailty and comorbidity. Younger patients will often prioritise treatment options based on increased life expectancy; older patients will often prioritise options based on quality of life [87]. These preferences and differing clinical needs are reflected in core care plans and patient care plans. This is especially true with chemotherapy. Chemotherapy is recommended only for stage III or high-risk stage II cancers; older patients with higher stage cancers will be less able to tolerate chemotherapy, and the increase in life expectancy will be minimal but with a huge cost in terms of side effects and time spent out of home and in-hospital for administration.

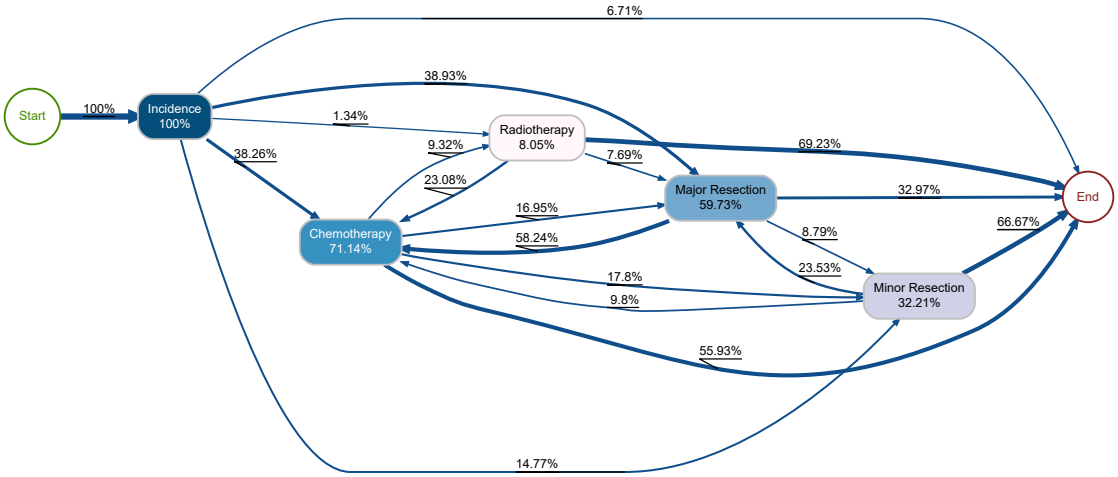


Figure 3.3: DFG representing treatment journeys for colon cancer patients aged 50 - 59 at incidence.

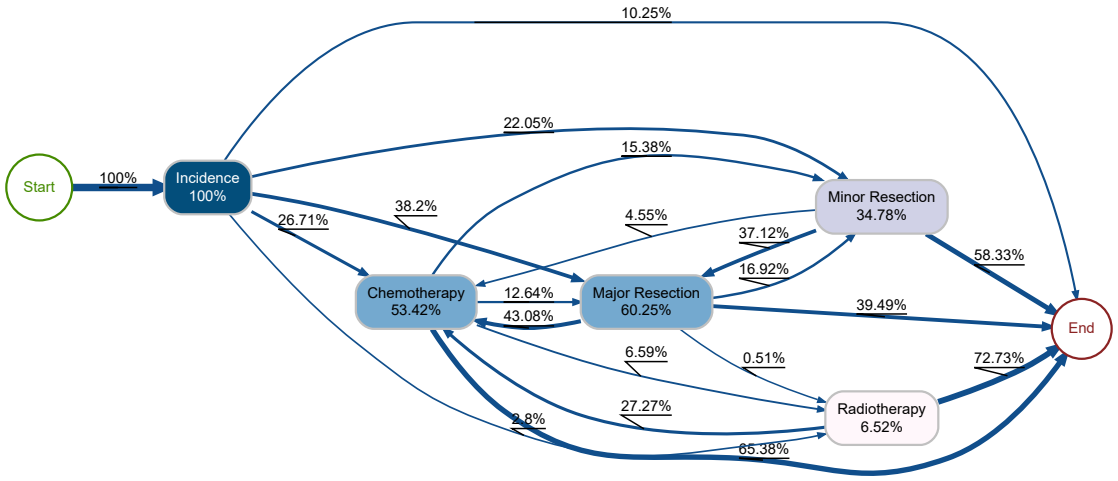


Figure 3.4: DFG representing treatment journeys for colon cancer patients 60 - 69 at incidence.

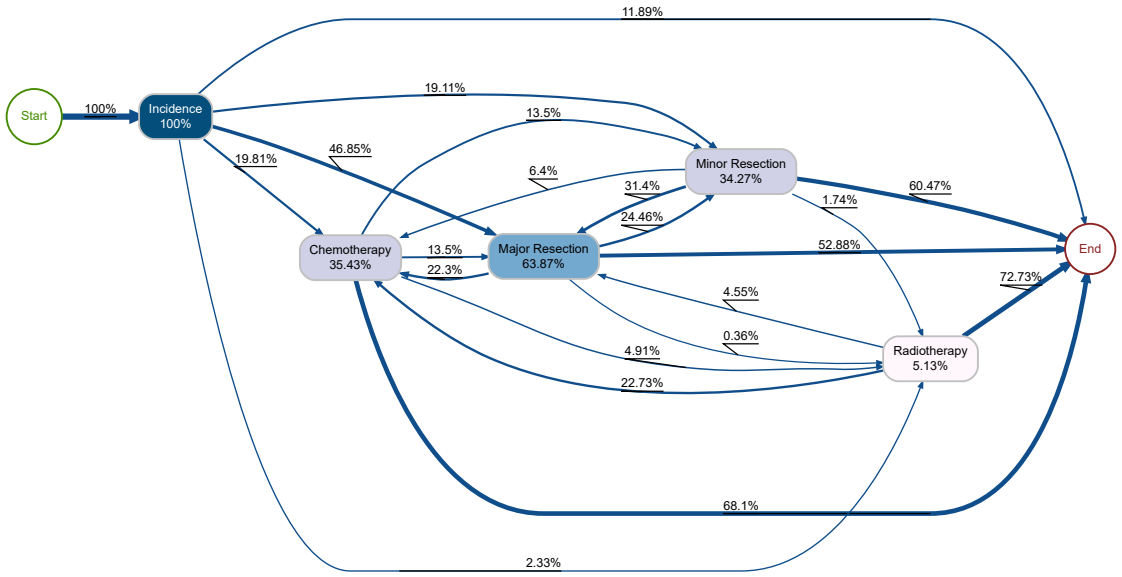


Figure 3.5: DFG representing treatment journeys for colon cancer patients aged 70 - 79 at incidence.

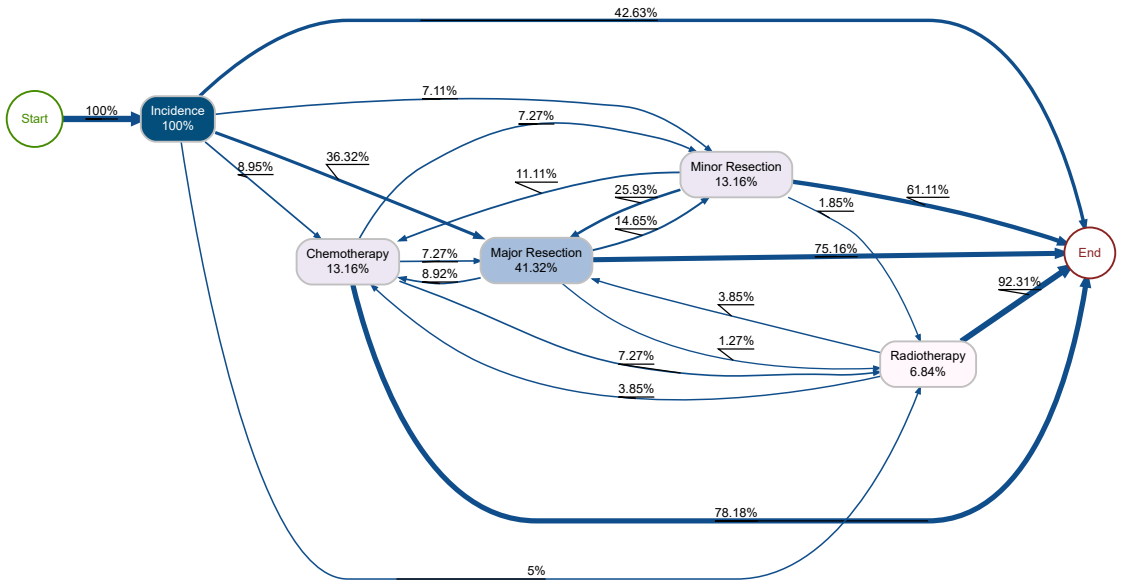


Figure 3.6: DFG representing treatment journeys for colon cancer patients aged over 80 at incidence.

3.3.4.3 Charlson Score

The patient population was then stratified using a common measure of comorbidity, the Charlson score, calculated using the *comorbidity* R package [88]. This score is a summary statistic that takes into account the presence of 17 different diseases associated with mortality risk. It is used in epidemiology to adjust survival models to make research comparable across patient populations. A higher score indicates the presence of more comorbidities, and therefore a higher risk of early mortality [89].

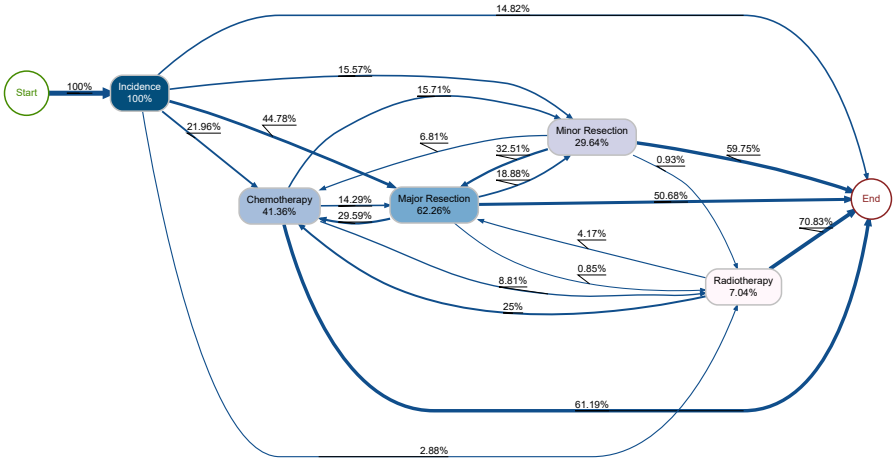
It is a very blunt measure that has many biases as a patient has to have been treated at a hospital for one of the relevant conditions; therefore, patients who are managed outside of the hospital will never have a relevant diagnosis recorded in their hospital care record. Many diseases that are actively managed in the community, such as diabetes, will not be recorded in a hospital care record, artificially decreasing a patient's Charlson score. 349 patients, or 25%, did not have any hospital records previous to their cancer diagnosis and therefore are marked with a Charlson score of 0, even if this does not reflect reality.

Due to the low patient numbers associated with a Charlson score of 3 (<20), only patients with a Charlson Score of 0, 1, and 2 are considered. Figure 3.7 shows each of these groups' directly-follows graph.

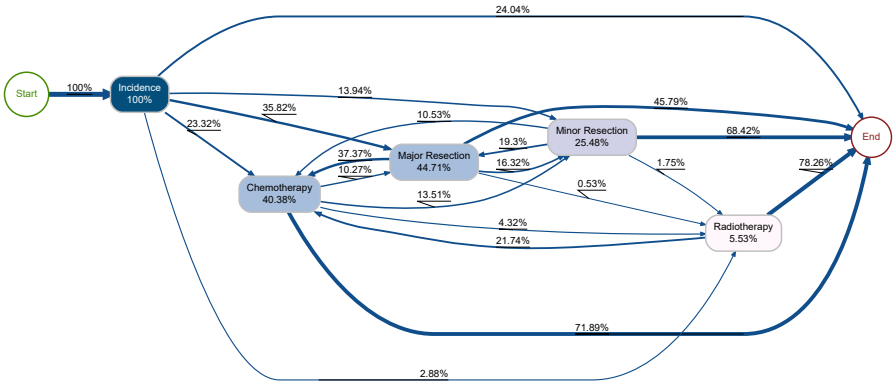
As expected, rates of non-treatment increase with Charlson score, going from 15% of patients with a Charlson Score of 0 (Figure 3.7a), to 24% with a score of 1 (Figure 3.7b), and finally 47% for those with a Charlson score of 2 (Figure 3.7c).

For individual treatments, the usage of major resections decreased the most, going from 62% in those with a score of 1, 45% with a score of 1, to 25% in those with a score of 2. These findings are expected clinically, as increased frailty will reduce a patient's ability to undergo intensive surgery, or even to undergo any form of treatment.

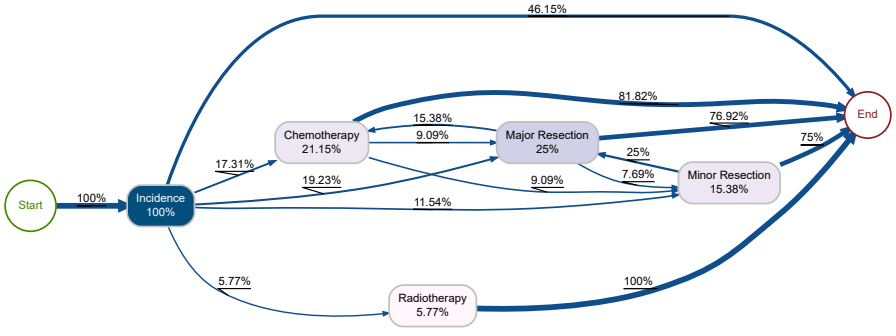
Chemotherapy rates are similar in patients with Charlson 0 and 1, at 41% and 40% respectively, but half to 21% in patients with a score of 2. As Charlson score is highly correlated with age– 65% of those with a Charlson Score of 2 are aged over 70– these journeys therefore should mirror the patterns seen in the age subcohorts.



(a) Charlson comorbidity score of 0



(b) Charlson comorbidity score of 1



(c) Charlson comorbidity score of 2

Figure 3.7: DFGs representing treatment journeys for colon cancer patients with different Charlson scores.

3.3.4.4 Socio-economic Factors

While it is clinically relevant to see variations in care on the basis of patient age or health state, variations associated with socio-economic background are less welcome and the subject of considerable concern, effort, and investment across public health systems [84].

To investigate this the patient population was stratified using the English indices of multiple deprivation (IMD), which ranks localities based upon income, employment, education and skills, health and disability, crime, barriers to housing and services, and the living environment. Patients are assigned an IMD based on their home post-code. The score used in this study comes from 2015 (IMD 2015), with the indices grouped into quintiles.

The DFGs show that based on IMD, patients at this NHS Trust are receiving similar rates of treatment. Of those patients living in the most deprived areas (the first quintile), 16.7% did not receive any treatment compared to 17.4% in the least deprived areas (the fifth quintile). The major resection rates were 52% and 56% respectively, and the chemotherapy rates were 52% and 41% as shown in Figure 3.8 and Figure 3.11. The difference is larger though when comparing the least deprived patients to patients living in second most deprived areas (second quintile, Figure 3.9): 33% have no treatment, with major resection rates of 44% and chemotherapy rates of 41%.

To investigate the differences further, Chi-squared tests were calculated to see if rates of no treatment, receiving chemotherapy, and receiving a major resection were statistically different between IMD groups. As the patient populations were small for IMDs 1 and 2 , they were combined to compare (shown in Figure 3.10) against IMD 5 for more comparable numbers and to reduce the risks associated with analysing small numbers. 26.8% of patients in IMDs 1 and 2 did not receive treatment, with major resection rates of 47% and chemotherapy rates of 45%. A Chi-squared test showed there is a significant difference in no treatment rates between IMD groups (X-square=5.9, p=0.02). There is also a significant difference in major resection and chemotherapy rates between groups (X-square=9.7, p=0.02).

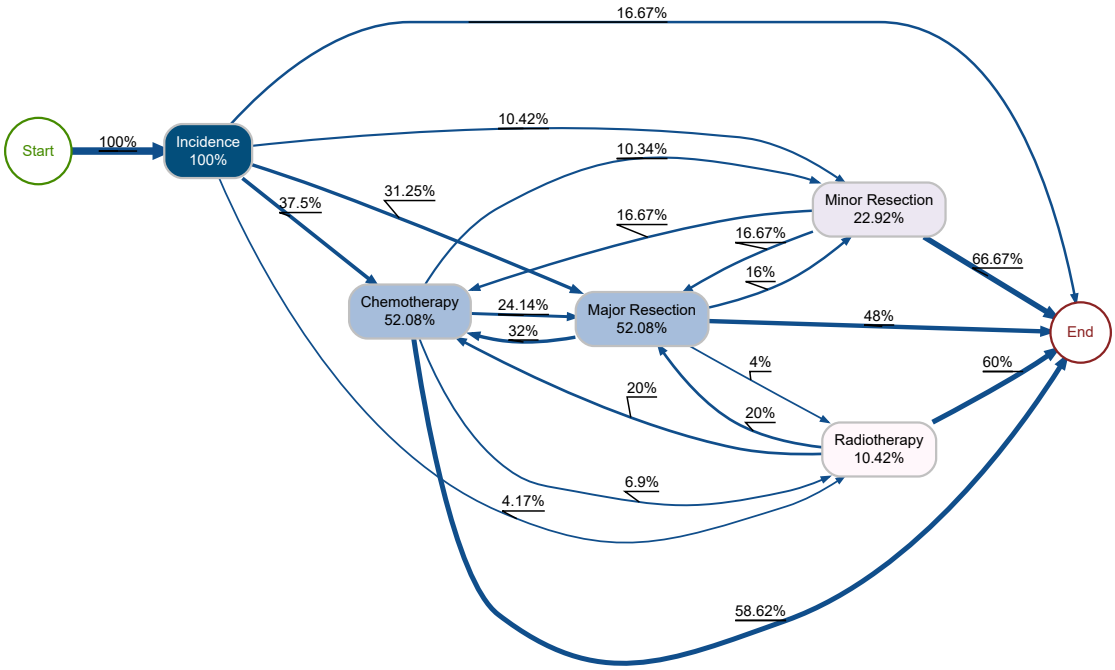


Figure 3.8: DFG representing treatment journeys for colon cancer patients living in a post-code with an IMD score of 1.

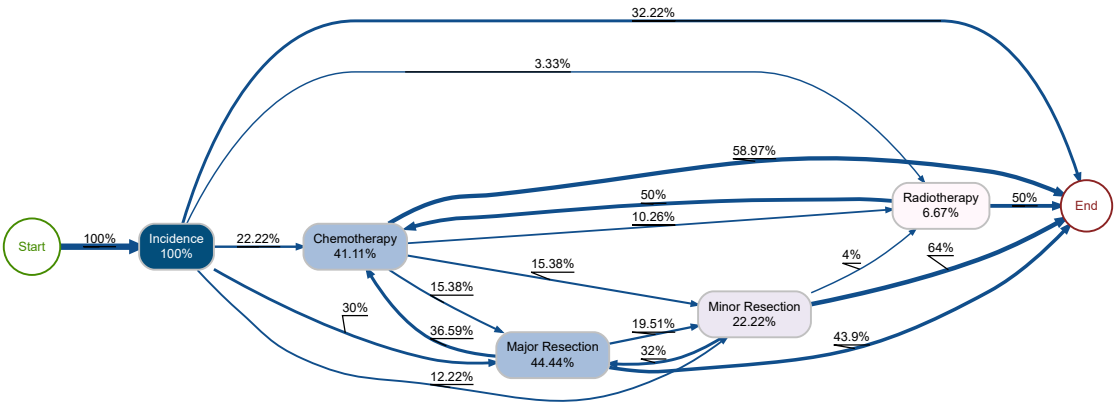


Figure 3.9: DFG representing treatment journeys for colon cancer patients living in a post-code with an IMD score of 2.

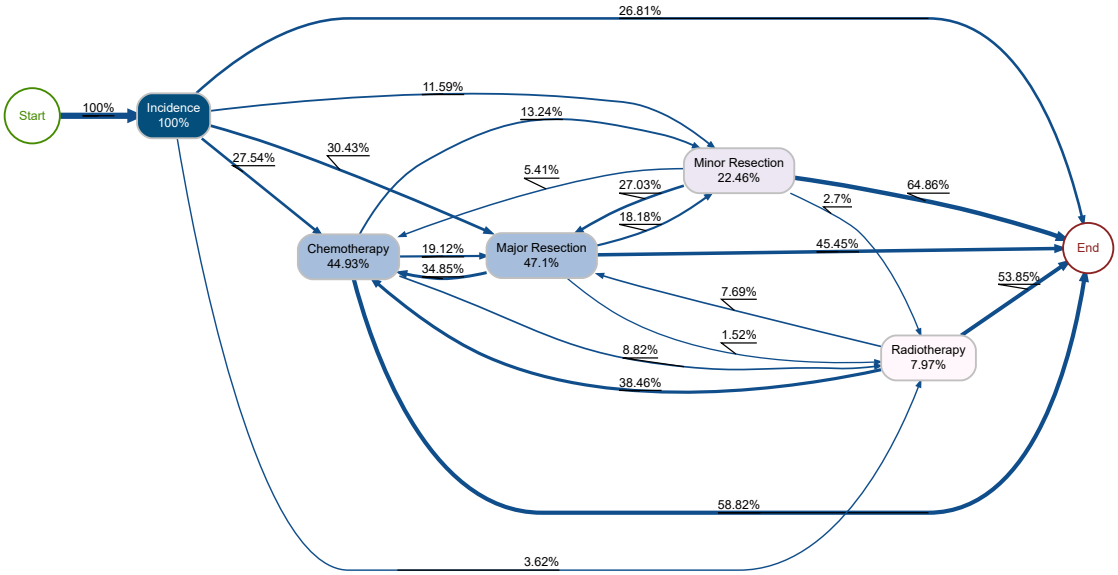


Figure 3.10: DFG representing treatment journeys for colon cancer patients living in a post-code with an IMD score of 1 or 2.

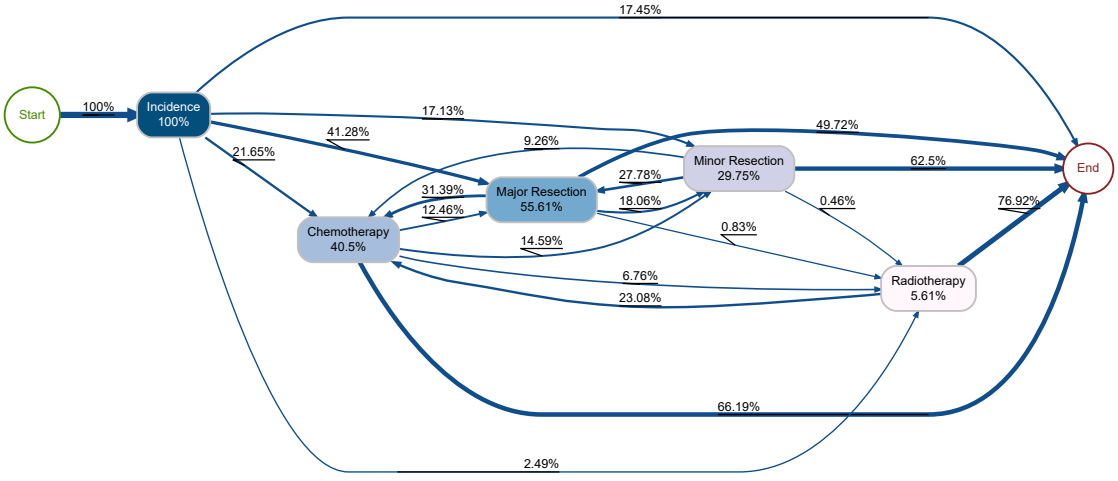


Figure 3.11: DFG representing treatment journeys for colon cancer patients living in a post-code with an IMD score of 5.

3.3.4.5 Time to Major Resection

NHS England national guidance states that patients with cancer should receive treatment within 62 days of being referred to a hospital for suspected cancer [62], with a national target of 85% of patients meeting the waiting time.

A subcohort of 580 patients was considered who received a major resection as their first treatment. 18% (102 patients) did not receive treatment within the 62 day of incidence; incidence was used as a proxy as referral date was unknown. All events between the incidence date and the date of resection were then isolated for these patients.

For 25% of patients there were consistent diagnostic imaging or endoscopic events between incidence and resection. Half of these had their major resection within 62 days of their last endoscopic event.

Another 25% had a X29.2 code, which means "Continuous intravenous infusion of therapeutic substance NEC" in-between incidence and resection. The OPCS clinical coding guidelines, *Chemotherapy Regimens Clinical Coding Standards and Guidance OPCS-4* [90] and Terminology and Classifications Delivery Service [36] state that OPCS code X29.2 should be used when a patient has been admitted for the sole purpose of administration of G-CSF, a growth hormone used to stimulate white blood cells. It is used in-combination with chemotherapy, to combat side-effects, or to treat patients with an infection. Over 80% of these patients also had a lung functional test or heart echocardiogram around the same time. This indicates that the delay in resection was potentially due to the frailty of the patient.

A few patients (9%) who had an incidence date in November subsequently had their major resections end of January or early February. This delay in treatment may be due to decreased capacity, due to staffing schedules and increased winter pressures, of the hospital during the Christmas period. These are examples of contextual information which is not coded into a patient's care record, but can affect their care pathway.

9% had another operational procedure, such as ablation of the lungs or a stoma-forming procedure, between their incidence and resection dates. These are clinically

valid reasons for missing the target, as procedures require rest in between. For the remaining 32%, I couldn't identify a reason for the delay as there was no coded information in their care record between the events of interest.

Therefore, 59% of cases had a clinically valid reason for delayed treatment when looking beyond the treatment events of interest, highlighting the importance of considering data points and mining for events beyond the initially identified codes of interest. This also shows that when using typical process mining conformance checking techniques, hard rules may result in over-penalising pathways which are inherently meant to be flexible and updated as a patient's health state changes. It is also important to consider other factors which may impact care delivery, such as holidays and hospital capacity fluctuations.

3.4 Importance of Data Provenance

When Section 3.3 was originally published in Chammas, Dwyer, Sallinger, *et al.* [51] the patient population was selected using approach 1: finding patients with an ICD-10 C18 code somewhere in their general care record. All the data came from the hospital's main inpatient and outpatient systems, Orbit, reflecting its feed into the Hospital Episode Statistics. For this chapter, the analysis was repeated using a patient population selected using approach 2, selecting those with a ICD-10 C18 code in the cancer management electronic patient record system, Infloflex. The change in patient population numbers is explored in Section 3.3.2.

Basing the patient population on the cancer management system restricts the patients to those with a confirmed, actively managed case of cancer versus relying on the diagnostic codes attached to procedures such as colonoscopes, which can indicate a suspected versus confirmed diagnosis. This restriction ensures that we are looking at actual treatment journeys, especially when looking at rates of no treatment, which had the biggest difference from the initial study results in Chammas, Dwyer, Sallinger, *et al.* [51]. However, this restriction does not allow the hospital to understand how many patients with colon cancer are not making it to MDT, for example, due to early death. Having a definitive source

of cancer diagnosis, such as the National Cancer Registry, linked to the hospital's care records would allow for this analysis.

Another example of how this difference in population definition changes the results is when looking at treatment journeys for patients with a Charlson Score of 0. In the original study, 82% of patients with a Charlson Score of 0 had no treatment events on their pathway– a finding that goes against the practice that patients with less comorbidity should have more treatment [51], [87]. Investigation into this population found that for majority of the patients, the C18 code was used to indicate "suspicion of malignancy" versus a confirmed diagnosis. The restriction of the patient population in this study resulted in a no treatment rate of only 15%, the lowest for all the groups.

Another difference in the data for this study is that the chemotherapy and radiotherapy events came from the treatment-specific EPR systems versus the general outpatient EPR systems. While there wasn't a large difference in the number of chemotherapy events between studies, the number of radiotherapy events did increase by about 3%.

3.5 Discussion

I developed epidemiology-informed process mining to address the recommendations routinely presented in the healthcare process mining literature. Most directly, looking at the process through the patient's eyes, involving domain experts at every stage, considering healthcare specificities during technique development, and developing a comprehensive methodology that takes these specificities into account [10], [13], [15], [47].

These recommendations and challenges reflect the importance of context and domain expertise: a specific question for a particular end-user provides a basis for selecting different kinds of source data and for deciding upon an appropriate classification of events. Domain expertise is needed to ensure that not only the question, but also the semantics of the data– its provenance and interpretation – are properly understood. An account of data preparation and analysis methods

is needed to determine the applicability and generalisability of any results: the epidemiology-informed process mining approach considers all of this.

Colon cancer patient journeys at the studied NHS Trust differed in patient cohorts defined by age and comorbidity score, and there were small differences across deprivation. By considering additional information in the care record, I could determine reasons for delays in patient treatment. For example, for the majority of patients whose treatment came later than the national guidelines, I was able to establish that the reason for the delay was clinically valid: a competing diagnosis resulting in increased frailty or a care management decision based upon diagnostic information. This shows how having knowledge of clinical codes is not enough to support analysis; additional, contextual information is required.

An important limitation of this method concerns the feasibility of obtaining high-quality contextual information at scale from EPRs. Each of the systems used within a hospital will have a different means of representing and accessing relevant data. While core demographic and treatment information may be retrieved from standard reports provided by these systems, data corresponding to intended care plans is difficult to extract and standardise. Additionally, proxies created to generate contextual information about the patient, such as the Charlson score, are blunt and can form biases, as patients need to have been treated at hospital for their pre-existing conditions in order for the diagnosis to be present in the care record. A lack of a diagnosis in the hospital care record does not mean the patient does not have that disease or is not being treated for it elsewhere. Data points that would have significantly benefited this analysis include the patient's cancer stage, their intended care plan, and the clinician's decisions for treatment. Without these data points, these findings cannot go beyond observational study as the context for treatment decision-making cannot be fully analysed.

Another limitation of the approach is the lack of concrete guidance for the presentation of findings and their interpretation at the end of each proposed iteration. This was not addressed in this chapter as the focus was to create a generalisable method, and methods for result interpretation will change based on the research

question and data available. The recently published Grisold, Aa, Franzoi, *et al.* [91] provides a framework for "context-based sense-making of process mining results". They suggest three context levels for sense making of results–process-immediate, organisation-internal, and organisation-external– alongside further dimensions, descriptions, and examples for each. The framework is compatible with epidemiology-informed process mining. Future work could look to standardise its application for the healthcare context, as the framework's initial application focus is on financial institutions. Future work will also seek to define more concretely how to embed statistical analysis in the approach to effectively determine if differences in processes between patient cohorts are meaningful (not due to chance), as well as provide an avenue to determine drivers of process variation.

Lastly, the analysis of patient journeys by stratifying patients by Charlson score and IMD should be standardised for age. This is because there can be a correlation between a patient's age and their Charlson score, and there can be varying effects of IMD by age. Without standardising for age, the analysis will not take these correlations into account and the results can be biased or show a relationship which is not there. The methods to do this appropriately will be apart of future work, resulting in practical guidance for process miners on how to handle for hidden correlations and other confounding in their statistical analysis of patient pathways.

The combination of process mining and epidemiology has the potential to accelerate healthcare transformation. The process mining method outlined in this chapter is informed by epidemiology. There are clear opportunities for epidemiology, in turn, to be informed by process mining. At present, a typical epidemiological study involves checking to see whether receiving a particular treatment is associated with an improvement in outcomes [16]. Using process mining, not only whether a particular treatment was received can be considered, but also when it was received and whether it came before or after other treatments; this will be considered further in Chapter 5.

"The model should emphasise the things relevant for a particular type of user." –van der Aalst, Adriansyah, de Medeiros, et al. *[11]*

4

Which Events are in a Patient's Journey?

Contents

4.1 Introduction 53
4.2 Preparing Healthcare Data 55
4.2.1 Event Selection in PM4H 56
4.2.2 Code-lists in Clinical Research 57
4.3 Ontology Based Code-list Generation 58
4.4 Study 1 60
4.4.1 Queries 60
4.4.2 Activity Abstraction 63
4.5 Study 2 69
4.5.1 Recall and Precision of Query Code-lists 71
4.5.2 Discovered Processes 73
4.5.3 Enriching Code-lists with Statistics 77
4.6 Discussion 79

4.1 Introduction

This chapter presents a new method for activity selection during event-log generation for process mining in healthcare. This method continues the thesis's goal of aligning process mining methods with those used in clinical research by drawing a parallel between activity selection, as performed in process mining, and code-list curation, as used in clinical research. It investigates the extent to which event-log generation,

including activity abstraction, can be automated by leveraging the medical ontologies and standardised clinical terminologies that care records are coded in, and shows how process models can be enhanced by structured domain knowledge. This chapter address challenges and recommendations 1, 3, and 5 as outlined in 3.1.

To investigate how ontologies can be used to automate event-log preparation, SNOMED's Expression Constraint Language (ECL) is queried to return lists of candidate procedure concepts related to a disease of interest using four different sets of constraints. The returned lists are then evaluated against their agreement with established, expert-curated code-lists for clinical research. Two real-world EHR datasets are used to demonstrate the importance and effects of code-lists and activity abstraction methods. Dataset A is EHR data from a major English regional NHS Foundation Trust, also used in Chapter 3. Dataset B is a publicly available EHR dataset from the USA called MIMIC [49].

Study 1 generates procedure candidate code-lists for colon cancer using four different ECL queries. The query derived SNOMED code-lists are then converted into OPCS codes using NHS Digital's classification mappings [92]. The query code-list is then compared to an expert curated code-list from the CORECT-R project [72]. Event aggregation and abstraction methods on both the ECL query code-lists and the reference code-list are evaluated by applying them to Dataset A and investigating their effects on Dataset A's colon cancer treatment process trace variants.

Study 2 expands on Study 1 by applying the ontological method to generate candidate code-lists for five different diseases: appendicitis, cataract, glaucoma, colorectal cancer, and lung cancer. Additionally, it uses a different clinical coding conversion method– converting the procedure candidate lists from SNOMED into OPCS and ICD using the Observational Health Data Sciences and Informatics' OMOP vocabulary mappings [93]. Event-logs and process maps are then generated and compared for colon cancer and lung cancer patients using Dataset A. Finally, the query code-lists are also evaluated using a proposed statistical method using Dataset B.

The work in Chapter 4, Study 1 was first presented at the International Workshop on Process-Oriented Data Science for Healthcare (PODS4H) at the International Conference on Process Mining (ICPM) in 2022, where it received a "Best Student Paper" award [94]. The paper was then invited for extension and submission to a special edition of the Journal of Intelligent Information Systems, which is under review, and forms Study 2. Both papers were co-written, and the methods co-developed, with fellow Health Data Science PhD candidate Owen Dwyer. Study 1 was entirely re-written with new figures produced by myself for this thesis. Study 2's figures were produced by Owen Dwyer with the analysis and discussion written by myself for this thesis.

4.2 Preparing Healthcare Data

Generating event-logs from care records requires translating the language of notes, images, test results, messages, letters, and codes into a language of abstract events. As this data has been collected for other, more immediate purposes discussed in 2.3, this translation may not be wholly obvious. Data interpretation often rests on understanding the mechanisms of data generation and having a clinical context that is often not recorded, or must be deduced through corroboration of other data elements [74]. Even when this translation is based largely upon the aggregation of clinical codes, this step may require a considerable amount of iteration and interaction with domain experts [13], [74], [95], [96]

The first step in generating event-logs is to determine which clinical activities should be mined from the care record. As a patient's care record will include every event related to their medical history, there needs to be filtering to ensure that the journey of interest is correctly mapped. Next, the abstraction level, or levels, of the clinical events must be selected. Both steps should be determined by the research question, which defines the patient journey of interest, why it is being mined, and the target audience of the results.

4.2.1 Event Selection in PM4H

When generating event-logs from patient records, several data quality issues and specificities of healthcare data and delivery must be considered. This includes understanding the "manifestation of process characteristics in event-logs" and the manifestation of the process in a healthcare setting [42], [97]. Filtering is a common data pre-processing method and can occur at the event and the trace level [76], [98]. It can happen statistically by searching for the most frequent, correlated, or even outlier, events. This risks missing rare but significant events in a patient's journey or selecting for very common but unrelated events to the clinical pathway of interest. Munoz-Gama, Martin, Fernandez-Llatas, *et al.* [15] even recommends that PM4H researchers "must go beyond simply filtering out infrequent behaviours from the event log".

A standard approach to generating event logs from a publicly available healthcare dataset has recently been proposed by Cremerius, Pufahl, Klessascheck, *et al.* [76]. In this approach, event filtering involves human curation with process mining researchers choosing procedures in a terminology based on those described in relevant clinical guidelines. This approach, however, is limited by the typically vague language of clinical guidelines. For example, the UK's guidance on colorectal cancer simply recommends that "surgery" be offered to patients: this does not specify a particular code or a clear rule for which specific events should be included in a study, and will be interpreted differently by different healthcare systems and individuals [65]. Guidelines will also differ by location, making comparing results difficult. There is therefore a need to explore the event selection stage in deeper detail, and establish clear and consistent rules for what is and is not included.

The usage of semantics and ontological reasoning to enhance process mining has been explored in the past. Alves de Medeiros and van der Aalst [99] describes semantic process mining as the explicit annotation of elements in a log with the concepts they represent, thereby making it possible to infer other relationships between them automatically; they even suggest that such ontologies might be useful in the log cleaning process.

Other approaches have used reasoning for abstraction, to group granular events into fewer categories for analysis, and for aggregation, to merge common event sequences into single events based on rules [74], [100]. Work has also highlighted the apparent gap between fine-grained events encoded in EHR data and the more abstract events described in treatment guidelines, with ontology-based abstraction proposed as a way to bridge this gap [101].

4.2.2 Code-lists in Clinical Research

In clinical research, data filtering often happens through domain expert curation of relevant clinical codes in the care record's clinical terminology called a *codelist* [102]. While this increases the reliability and relevance of event selection, it is resource and time-intensive, especially when there might be thousands of possible events in the care record. It must also be reproduced for different datasets using different terminologies and for each disease of interest.

An automated, or at least semi-automated, code-list curation process has been attempted before: for example, Watson, Nicholson, Hamilton, *et al.* [102] framework selects codes based on whether their text descriptions contain specific search terms. Here, the authors recommend that this step should be done cautiously, retaining codes unless a clear consensus exists to reject them.

The usage of semantics and reasoning to assist in the preparation of code-lists has also been explored before. Elkheder, Gonzalez-Izquierdo, Qummer Ul Arfeen, *et al.* [103] demonstrates how reasoning over SNOMED relationships can produce diagnostic code-lists, and that these code-lists produce similar results to their handcrafted equivalents when used to select patients from a larger cohort. A natural evolution of this work is, therefore, to ask whether such methods would also be effective for procedure codes and whether the resulting process models would resemble models created using expert-curated code-lists.

4.3 Ontology Based Code-list Generation

To utilise the medical knowledge stored in SNOMED's ontology, SNOMED's ECL was queried to return concepts of interest from the SNOMED ontology [104], [105]. The use of semantic relationships in SNOMED allows for logical inferences to be made between concepts, which can be exploited to automatically discover procedure codes that might be relevant to a particular disease or symptom.

Four ECL queries were tested by formalising research questions into queries, and then iterating on these queries. The ECL queries were run on SNOMED CT's (web) browser. The returned SNOMED code-lists were then converted to either OPCS or ICD codes for analysis against already published, reference code-lists [92]. The queries were run against the SNOMED-CT UK Edition, 2024-02-14 release.

The first question was "which procedures occur on a site that is a *possible* finding site for our target diagnosis?", with the corresponding ECL query shown in Listing 4.1. A query starts with the SNOMED numerical code for the concept of interest, followed by the name of the code in the SNOMED Code System Concept: for example, 71388002 represents the code system concept "procedure" while 405813007 represents the code system concept "procedure site - direct" [105].

```
< 71388002|Procedure| :
  405813007|Procedure site - Direct|
  =<< (* : R 363698007|Finding site|  = \pmb{Diagnosis})
```

Listing 4.1: "Which concepts are procedures, and occur on a site that is a possible finding site for our target diagnosis?"

To test Query 1, the query was written to return procedures occurring on a possible finding site of target diagnosis colon cancer (e.g. procedures with a site of colon). Query 1 returned 644 codes, many with only tangentially relevance to colon cancer. For example, a number of codes describing *oesophagectomy*, or the surgical removal of the oesophagus, were returned because this procedure can involve replacing the oesophagus with material from the colon. This results in the colon being tagged as a procedure site for this concept in the SNOMED ontology.

While the colon is a listed secondary organ procedure site for a *oesophagectomy*, it is a procedure unlikely to be related to colon cancer.

Therefore, the additional constraint that the procedure must have one and only one site was added to the query to exclude tangential concepts, creating Query 2 in Listing 4.2. Query 2 reduced the number of codes returned by almost half to 333 because of the removal of many procedures which had non-colon primary sites, such as the spleen or oesophagus.

```
< 71388002|Procedure| :
  ( 405813007|Procedure site - Direct| =
    << (* : R 363698007|Finding site|  =<< Diagnosis) )
  AND ( [1..1] 405813007|Procedure site - Direct| = *)
```

Listing 4.2: "Which concepts are procedures, occur on a site that is a possible finding site for our target diagnosis, and have one and exactly one procedure site?"

Query 1 also returned all codes related to metastatic sites of colon cancer. For example *malignant neoplasm of colon* had child concepts including *malignant neoplasm of recto-sigmoid junction metastatic to brain*, resulting in the brain being automatically introduced as a possible procedure site of interest. This then results in a number of codes relating to the brain being returned in the code-list.

To prevent this from happening, Query 3, shown in Listing 4.3, was written requesting only procedures that involve a named anatomical site and its child concepts, versus requesting any possible site of the target disease. Query 3 returned 303 codes and restricted the number of OPCS Y-codes returned, which represent "subsidiary classification of methods of operation" and therefore are site agnostic.

```
< 71388002|Procedure| :
  405813007|Procedure site - Direct|
  =<< (* : R 363698007|Finding site|  =<< Site )
```

Listing 4.3: "Which concepts are procedures, and occur on a specified site?"

Query 4, shown in Listing 4.4, combines Queries 2 and 3 to constrain the output to procedures with exactly one procedure site *and* a procedure site that is a specific named site, or a descendant thereof. This returned the smallest code-list with 193

concepts.

```
< 71388002|Procedure| :
    ( 405813007|Procedure site - Direct| =<< Site )
    AND ( [1..1] 405813007|Procedure site - Direct| = *)
```

Listing 4.4: "Which concepts are procedures, occur on a specified site, and have one and exactly one procedure site?"

Radiotherapy and chemotherapy treatments for cancer are not site specific, but highly relevant to studying cancer treatment. Therefore, an additional query was developed to request the relevant chemotherapy and radiotherapy codes, shown in Listing 4.5, which is used alongside Queries 1 - 4:

```
... OR << 367336001 | Chemotherapy (procedure) |
    OR << 108290001 | Radiation oncology AND/OR radiotherapy
                      (procedure) |
```

Listing 4.5: "Which concepts are chemotherapy procedures and radiation oncology and/or radiotherapy procedures?"

4.4 Study 1

4.4.1 Queries

The ECL derived code-lists were compared to a reference, domain expert curated code-list of colorectal cancer treatments, specifically the data coding catalogue from CORECT-R, to evaluate performance and benchmark the usability of logical reasoning for activity selection.

CORECT-R's code-list maps 197 different OPCS codes to five different treatment categories: major resection, minor resection, stoma, stent, and bypass. The list was extended for this thesis in consultation with domain experts to include additional clinical codes for diagnostic tests, chemotherapy, radiotherapy, and surgical procedures related to the maintenance of treatments, amounting to a total of 497 OPCS codes. This extended code-list is made publicly available on GitHub for other researchers to use [85].

The code-list returned by each ECL query was evaluated according to precision, the proportion of returned codes that were in the reference code-list, and recall, the proportion of codes in the reference code-list found in the returned code-list.

$$Precision = \frac{\text{Sum of returned codes in reference list}}{\text{length of reference code-list}}$$

$$Recall = \frac{\text{Sum of reference codes in returned code-list}}{\text{length of returned code-list}}$$

To allow for comparison the SNOMED codes were converted to OPCS codes to match the reference code-list using the NHS Digital's mapping files [92]. These mapping files are intended to assist clinical coders in converting SNOMED codes into OPCS for billing purposes. In the mapping files, many SNOMED concepts map to multiple OPCS codes. In these cases, clinical coders are instructed to either choose one or are expected to use all. For the purposes of this study, which is to generate a code-list with events that could be related to a disease versus definitely have to be, it is assumed that any OPCS code that is mapped in any way to a SNOMED concept could plausibly be related to it and is therefore retained.

4.4.1.1 Recall and Precision

Table 4.1 shows the precision and recall results of the four ECL queries against the original and extended CORECT-R code-lists. A consistent pattern amongst all the queries is that increasing the constraints on the query can drastically increase precision (more returned codes match the reference codes), but at a cost of lower recall (less reference codes are returned by the query).

Query	Precision		Recall	
	original	*extended*	*original*	*extended*
1	25.78	40.06	**84.69**	**57.59**
2	34.26	52.13	82.14	54.69
3	35.37	52.16	35.37	45.76
4	**50.62**	**73.44**	50.62	39.51

Table 4.1: ECL query code-list precision and recall against the original and extended CORECT-R code-lists

For the original CORECT-R code-list, Query 1 had the highest recall, retrieving 85% of the original CORECT-R codes. It also had the lowest precision rate of 26%, as it returned many codes not found in the original CORECT-R list. This is due to Query 1 being the least restrictive query resulting in a long, but non-specific, procedure code-list.

Of the original CORECT-R codes not returned by Query 1, 55% were OPCS codes ending in 8 or 9 (such as H15.8). These are catch-all codes representing "other" or "unspecified" variations of the parent code; H14 means *exteriorisation of caecum* while H14.8 means *other specified (exteriorisation of caecum)*, which have no direct mappings to the SNOMED ontology. The rest were Y-codes relating to stenting procedures which have no mapping to SNOMED concepts in the NHS mapping tables.

Query 4, the most restrictive query and the shortest code-list, conversely had the highest precision rate at 51%, meaning that 51% of the returned codes were in the original CORECT-R code-list. However, it had the third highest recall rate with only 51% of the codes in the original CORECT-R code-list being returned.

In the extended code-list results, precision rates are naturally higher and recall rates lower than in the original list since more of the returned codes are in the reference list, but a smaller proportion of the entire list is returned. Query 1 still had the highest recall of all the queries, but at a reduced rate of 58%, and still had lowest precision at 40%. Query 4 still had the highest precision at 73% but now the lowest recall at 40%.

Figure 4.1 shows the recall of each query further broken down by clinical event type. Codes representing blood tests, diagnostic testing and patient assessment were never returned by the SNOMED ECL queries. This is because they are not disease specific and therefore do not have a marked target site in the ontology, which is what the queries request against. Similarly, endoscopy of the upper gastrointestinal tract was not returned as it is tagged with a different procedure site in the SNOMED ontology than the colon, but it can be found on colon cancer

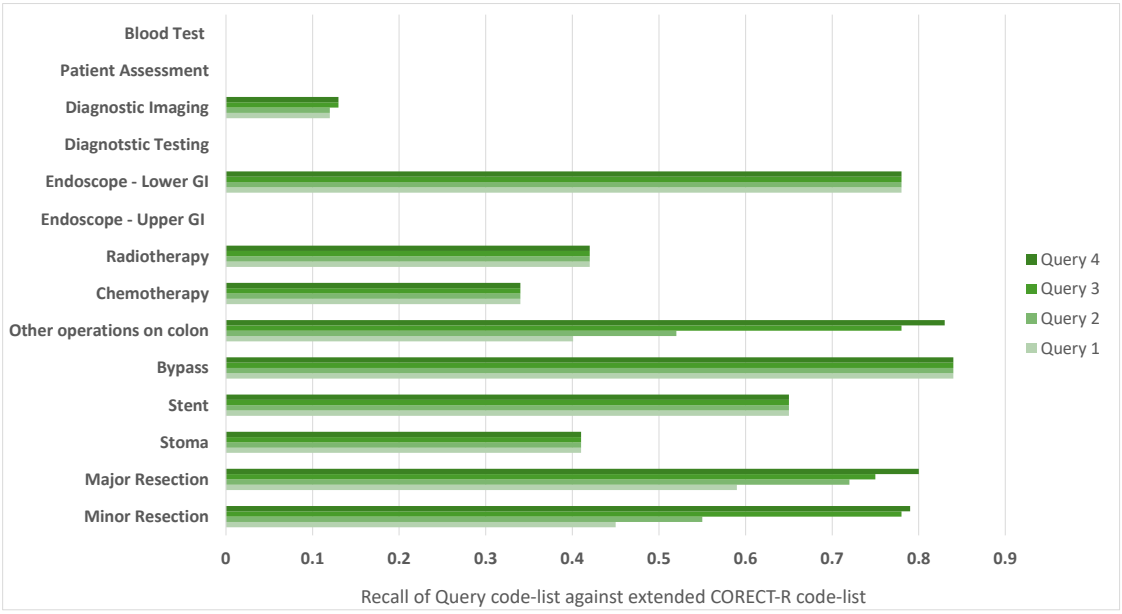


Figure 4.1: Recall of Query 1 - 4 compared to the extended CORECT reference set by clinical event type.

patients' route to diagnosis if they have unspecific symptoms, which is why it is in the extended CORECT-R code-list.

Recall was highest in surgical codes, such as major resection, minor resection and other operations on colon, as they are thoroughly annotated procedures with direct sites marked as colon in the SNOMED ontology. They were also the only topics to be notably affected by the difference in constraints between Queries 1 - 4 due to changes in the semantic relations being queried.

All other topics had the same recall between Queries; this is probably due to the specificity of procedure site for those codes. Chemotherapy and radiotherapy recall does not change between queries as they are returned by the same Query, shown in Listing 4.5 and appended to the Query 1-4's returned code-list.

4.4.2 Activity Abstraction

The Queries generate extremely long code-lists, which, if mined directly, will result in overly complex and uninterpretable "spaghetti process models." Therefore, methods are needed to abstract, or group, codes into meaningful and interpretable activities that can then be used to mine for events and generate processes.

Three techniques were used to investigate how ontology structure might aid in event aggregation or abstraction: according to the data's source clinical coding structure, according to semantic relationships in SNOMED, and finally, according to the text descriptions of clinical concepts. This study aims to show that the method for abstraction is just as important as the method for code-list generation in affecting process mining results.

4.4.2.1 Methods

The first, and simplest, approach uses the structure of the data's clinical coding terminology. A code in the OPCS terminology consists of four characters [36]. For example, the code H01.1 represents *Emergency excision of abnormal appendix and drainage*, however further qualified, H01.2 represents *emergency excision of abnormal appendix not elsewhere classified*, and H01.3 represents *emergency excision of normal appendix*. Collectively, all of these codes fall under the chapter H01 *emergency excision of appendix,* which itself falls under Chapter H *lower digestive tract*. Therefore, it is possible to aggregate codes by grouping them according to these chapters and subchapters.

Alternatively, a similar approach can be used within the SNOMED terminology, even before translating into another clinical coding terminology. SNOMED concepts are semantically annotated with additional relations, for example, procedures can have method or approach relationships. These relationships can therefore be used to infer group relationships, for example by categorising any concept with method imaging as "imaging", or anything with the methods repair or excision as "surgery". By examining the most common methods relationships that exist within the code-list, a set of maps can be manually determined that take advantage of this structured knowledge.

Finally, the text descriptions of each code can be used for aggregation, according to certain rules and conditions. This is the most manual and intensive approach, and less intuitive since it does not rely on existing ontological structure, but has the key advantage that it can be specified and developed by researchers according

to their needs, and categories can be designed to match existing categories from comparable code-lists or relevant national guidelines.

4.4.2.2 Method Comparison

Aggregating codes using the OPCS system's structure proved challenging. Top-level OPCS chapters are arranged by site (with the exception of a diagnostic imaging, testing, and rehabilitation and a miscellaneous chapter), which means that in practice, almost every code in a code-list related to a particular disease will come from one category. For example, Query 1 applied to colon cancer yielded 257 codes from the H (lower digestive system) chapter, 24 codes from the X (miscellaneous operations) chapter, and fewer than 20 codes from every other chapter.

Moving down a level to three-character OPCS codes, creates the opposite problem: there are far too many subcategories for this information to be useful. In the example of colon cancer, Query 1, returns 83 three-character concepts, each containing at most 10 codes, which is far too many to create an interpretable process model. This means that the OPCS code-system's structure in and of itself is not a useful framework for event aggregation into higher abstraction levels that still remain informative.

At the SNOMED level, the main barrier to useful aggregation was that many OPCS concepts had multiple possible categories, in most cases because the source SNOMED concept had multiple relationships, but occasionally because one OPCS concept could be mapped to multiple SNOMED concepts. This prevented the usage of the SNOMED relationship ontology for this study as the data itself was not coded in SNOMED.

A major barrier for using the structure of a clinical coding system for activity abstraction is that the categories did not clearly align with those from the reference code-lists, meaning that any resulting process models could not be directly compared.

For this reason, this study focuses on the third proposed aggregation method: text-based matching. A set of categories equivalent to those in the reference lists was defined and codes were matched based on a keyword search: for example,

any concept that contained 'resection', 'excision', 'lesion' or 'exenteration' in its description was mapped to 'resection'. These keywords were chosen based on patterns identified from manually examining the reference code-list to ensure close alignment. Codes were also filtered to only include those from OPCS Chapters for the target site of interest, and any additional non-site specific chapters such as U (diagnostic imaging, testing, and rehabilitation) and X (miscellaneous operations). Any Y or Z codes were removed as they are subsidiary classifications and should not be coded as a primary procedure by NHS clinical classification standards.

4.4.2.3 Effects on Trace Variants

To investigate the effects of a text-based matching activity aggregation method, Query 1's code-lists and the extended CORECT-R reference code-list were applied to Dataset A to filter the data and then return event-logs. Dataset A contains EHR data from a regional NHS Foundation Trust. Patients with a colon cancer ICD code (C18) somewhere in their records were included, resulting in a cohort of 4,127 patients. All inpatient and outpatient events after the first occurrence of the C18 code were retained and then filtered using either Query 1's code-list or the extended CORECT-R reference code-list.

27 activities were defined and matched in Query 1's code-list using the text-based matching method; 11 colon cancer specific procedures and 16 chapter headings. For example, chemotherapy is colon cancer specific procedure whereas other_A is for codes that came from the OPCS Chapter A section. 61% the 528 Query 1 codes were colon cancer specific while 39% were from unrelated Chapters.

In Dataset A, 172 clinical codes were found to match only 22 out of 27 reference activity categories. The codes dropped were non-site of interest chapter events. However, 511 instance of non-site specific chapter codes were returned for 479 unique patients.

10 event names are defined and matched in Query 1's code-list using the extended CORECT-R data mapping. 2 event names from the extended CORECT-R data mapping are not found in Query 1's code-list: these are blood test and patient

assessment. In the data, 177 clinical codes were found to match to all of the 10 filtered event categories.

Figure 4.2 shows the top 10 trace variants for Dataset A using Query 1's code-list with text-based abstraction, while Figure 4.3 shows the top 10 trace variants when using the extended CORECT-R abstraction table.



Figure 4.2: Top 10 trace variants for Dataset A using Query 1's code-list and text-based abstraction mappings. Columns 1-4 show the activity label in that step of the process. Column 5 shows the percentage of traces that row represents. Column 6 shows the number of patients with that trace. Column 7 shows the cumulative frequency of traces for that row.

The trace variants look extremely similar between the Figure 4.2 and 4.3. The differences lies in the granularity of the resection activity between the CORECT-R mapping and the text-based mapping. The text-based mapping groups together all resection events, whereas the CORECT-R mapping has two separate resection events: major and minor. Therefore, the second trace variant in Figure 4.2 is composed of Figure 4.3's second and sixth trace variant. Interestingly, there is a difference in the number of traces for the traces that are the same, for example "incidence, endoscopy" have less patients with that trace when using the text-based abstraction method than when using the CORECT-R abstraction method. The

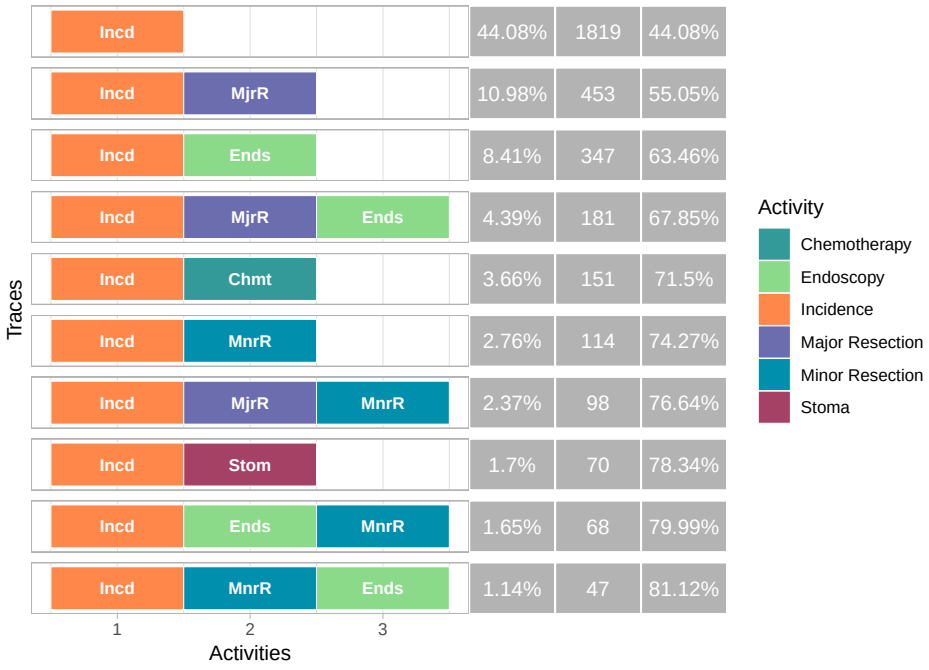


Figure 4.3: Top 10 trace variants for Dataset A using Query 1's code-list with extended CORECT-R abstraction mappings. Columns 1-4 show the activity label in that step of the process. Column 5 shows the percentage of traces that row represents. Column 6 shows the number of patients with that trace. Column 7 shows the cumulative frequency of traces for that row.

opposite is true however for the trace "incidence, stoma". One reason for this is maybe that when there is more patient's with that trace, there may be more clinical codes that map to that event, allowing for more instances of that event to be discovered in the care record.

What abstraction mapping you use, and when, is really important. By applying the CORECT-R mappings directly to the raw data, all of the abstract mappings listed in the mapping file will be returned as it selects for both the clinical codes and how to group them (Figure 4.4). If the CORECT-R mapping is placed over data which is first filtered using the Query 1 code-list (Figure 4.3), the resulting abstracted event-log will be different as the code-lists are different. This is shown directly by the fact that in Figure 4.3 there is no the inclusion of the event 'diagnostic imaging/testing', whereas in Figure 4.3 there is not. This is because the Query 1 code-list does not contain any codes related to diagnostics or imaging, so therefore it does not select for these activities.



Figure 4.4: Top 10 trace variants for Dataset A using the extended CORECT-R code-list and abstraction mapping.

4.5 Study 2

In Study 2, Query's 1 - 5 from Section 4.3 were used to derive code-lists to be compared against five different published code-lists covering five different diseases from three different sources, described in Table 4.2. This was to further validate the effectiveness of the ontology-derived code-list method, outlined in Section 4.3, on other disease groups.

Codelists A, B, and C, which represent appendicitis, cataract, and glaucoma, come from the HDRUK Phenotype Library [106]. The HDRUK Phenotype Library

Dataset	Disease	Number of Codes	Source
A	Appendicitis	14	HDRUK (PH106 / 212) [106]
B	Cataract	32	HDRUK (PH131 / 262) [106]
C	Glaucoma	21	HDRUK (PH176 / 352) [106]
D	Lung cancer	35	Consultation with domain experts
E	Colorectal cancer	497	CORECT-R [72] + extension

Table 4.2: Descriptions of the reference code-lists used in Study 2, including the disease, number of codes in the list, and the source including HDRUK phenotype library IDs and version numbers.

contains code-lists which define a disease, life-style risk factor, or biomarker. Therefore, the focus is not on procedural codes, which is the focus of this study. Codelists A,B, and C were chosen from the relatively small number of code-lists which contain procedural codes, and priority was given to longer code-lists in-order to produce more meaningful precision and recall statistics (for example, the diverticular disease code-list only provides one OPCS code, which would make evaluation meaningless).

Codelist D contains procedures related to lung cancer procedures and was constructed by consultation with domain experts; it is available at Chammas [85]. Codelist E is the original CORECT-R code-list published by the UK's Colorectal Intelligence Hub, also used in Study 1 [72].

These code-lists come from different sources and serve different purposes: A, B, and C act as minimum thresholds to determine if a patient should be considered to have a particular disease, life-style risk factor, or biomarker, whilst D and E are closer to exhaustive lists detailing all possible procedures for a disease in order to aid researchers studying the specific treatment processes. It is likely that code-lists D and E are produced to requirements that more closely resemble those needed for treatment process mining, but this range of different sources is deliberately included to consider how well this approach generalises to different scenarios.

The output of each ECL query tuned for each disease is evaluated according to precision– the proportion of returned codes that were in the reference code-list– and recall– the proportion of codes in the reference code-list found in the returned code-list.

Dataset A was used to create and compare process models using the SNOMED ECL query method and the corresponding reference Codelist to investigate treatments of lung cancer (ICD-10 codes C33*, C34*), requiring Codelist D, and colon cancer (ICD-10 code C18*), requiring Codelist E.

The third evaluation technique was a statistical method to determine the distribution of codes in real-world data to see if any codes were missed by the ontological approach. Dataset B, or MIMIC-IV, was used here. The colon and

lung cancer patients in MIMIC were divided into a case cohort of patients with the target diagnosis and a control cohort of those without. Through statistical testing with Fisher's exact test, codes with a statistically significant difference in frequency between the two cohorts, and therefore some association with a diagnosis of colon or lung cancer, were identified. The diagnosis correlated events were then compared to the ECL query code-lists to identify codes that have no clear logical link to the target disease, but are still associated with a diagnosis according to the data.

4.5.1 Recall and Precision of Query Code-lists

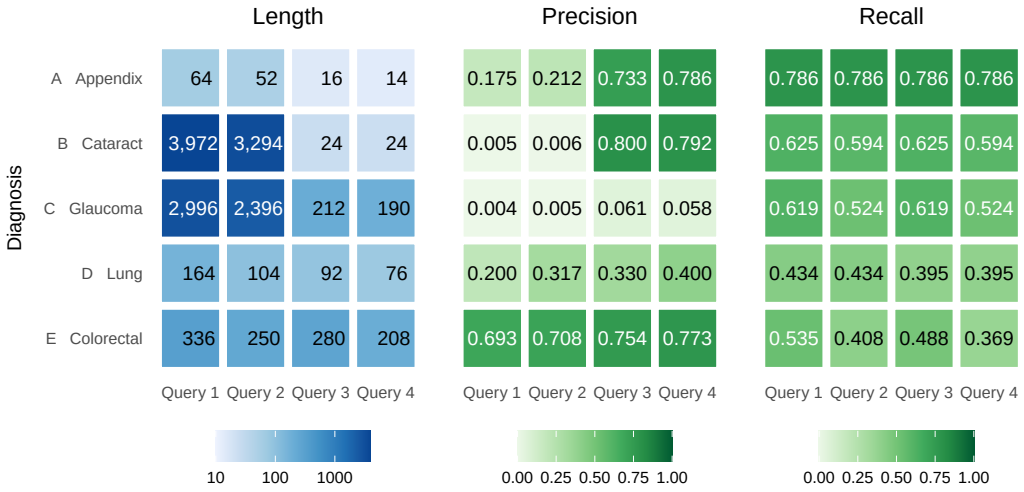


Figure 4.5: Length (number of codes in the list), precision and recall of SNOMED code-lists resulting from ECL query formals 1-4 against five different disease reference code-lists.

Figure 4.5 compares the results of Queries 1 through 4 for the five different disease scenarios. OPCS codes in chapters Y and Z (supplementary codes that describe additional information such as sites or approaches) are excluded from calculations since as a rule they never appear in the primary position in care records.

In terms of raw code-list size, moving from the more general Query 1 to the more restrictive Query 4 consistently reduces the number of codes returned, as expected. However, cataract and glaucoma produced exceptionally large result code-lists for Queries 1 and 2. This occurs due to the existence of concepts in SNOMED that represent multiple diagnoses, for example the concept *glaucoma and sleep apnea*

means that both *structure of the eye* and *structure of the respiratory system* are both introduced as acceptable procedure sites, so all relevant procedures to the eye and respiratory system are added accordingly. This issue is significantly improved by the introduction of the more narrow site definition in Queries 3 and 4.

Precision was generally low. This is most visible again in cataract and glaucoma, where the exceptionally large code-list size returned meant that inevitably a very small number of those codes would be relevant. However, it is also noticeable that in the case of the cataract code-list, and to a lesser extent in appendicitis, Queries 3 and 4 represented a significant jump in precision, suggesting that the move from querying for a specific disease to a single named procedure site eliminated a large number of irrelevant concepts. Colon cancer is the notable outlier in terms of precision, likely owing to the significantly larger size of the reference code-list compared to the others.

Recall, by contrast, was relatively high across all scenarios tested. It was the same across all queries for appendicitis, indicating that the changes to constraints did not affect the retrieval of matching reference codes. In every other case, recall was lower in Query 4 than in 1, indicating that increasing constraints—-while effective at reducing irrelevant codes as evidenced by the levels of precision—also removed some relevant codes.

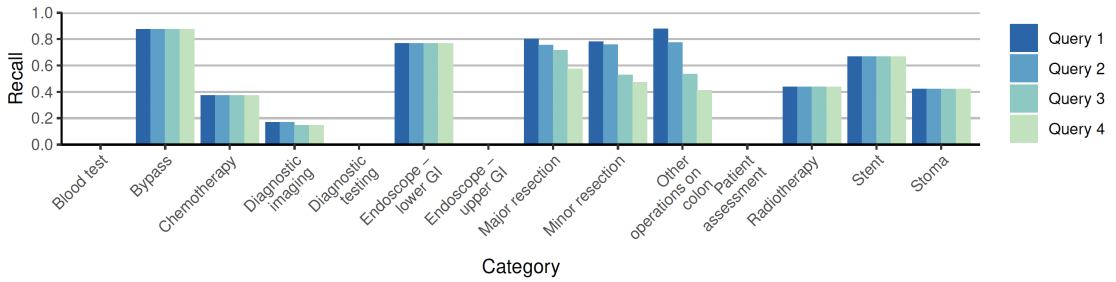


Figure 4.6: Recall of queries by category compared to the extended CORECT reference set

Figure 4.6 shows the recall of each colon cancer query, further broken down by the target concepts' categories according to the reference codelist. Concepts representing blood tests, diagnostic imaging and testing and patient assessment were never returned, presumably because they are not disease specific and therefore do not

have a clear site. Similarly, endoscopy of the upper gastrointestinal tract was never returned as it does not have a relevant procedure site to the colon, but it can be found on colon cancer patients' route to diagnosis if they have unspecific symptoms, which is why it is in the reference code-list. Recall was highest in surgical topics, since they were well annotated with clear procedure sites in the ontology. The differences in constraints between Queries 1-4 only had a noticeable effect on these codes.

4.5.2 Discovered Processes

To investigate the usefulness and effects of these code-lists on process discovery, directly-follows graphs for colon and lung cancer patients from Dataset A were generated using Query 1's generated code-list and reference Codelists E and D (with text-based abstraction for event names).

Directly follows graphs show the flow of cases (patients) between activities. The nodes are each activity, with the value inside showing the percentage of patients which receive that treatment at least once. The edges show how many patients move from that activity to the following activity.

4.5.2.1 Colon cancer

The journeys derived using Codelist E (Figure 4.7) had higher rates of treatment than using the ontology derived code-list (Figure 4.8). This is best reflected in the fact that 33% of patients in Figure 4.7 did not receive treatment while 60% did not receive treatment in Figure 4.8. As very few diagnostic imaging and test codes were selected in the Query 1 code-list, only 0.17% of patients had a diagnostic imaging event in Figure 4.8 whereas 37% had this event at least once in their journey using reference Codelist E. A large difference of event occurrence was also found in chemotherapy, with 8% of patients having a chemotherapy event in Figure 4.8, versus 21% in Figure 4.7.

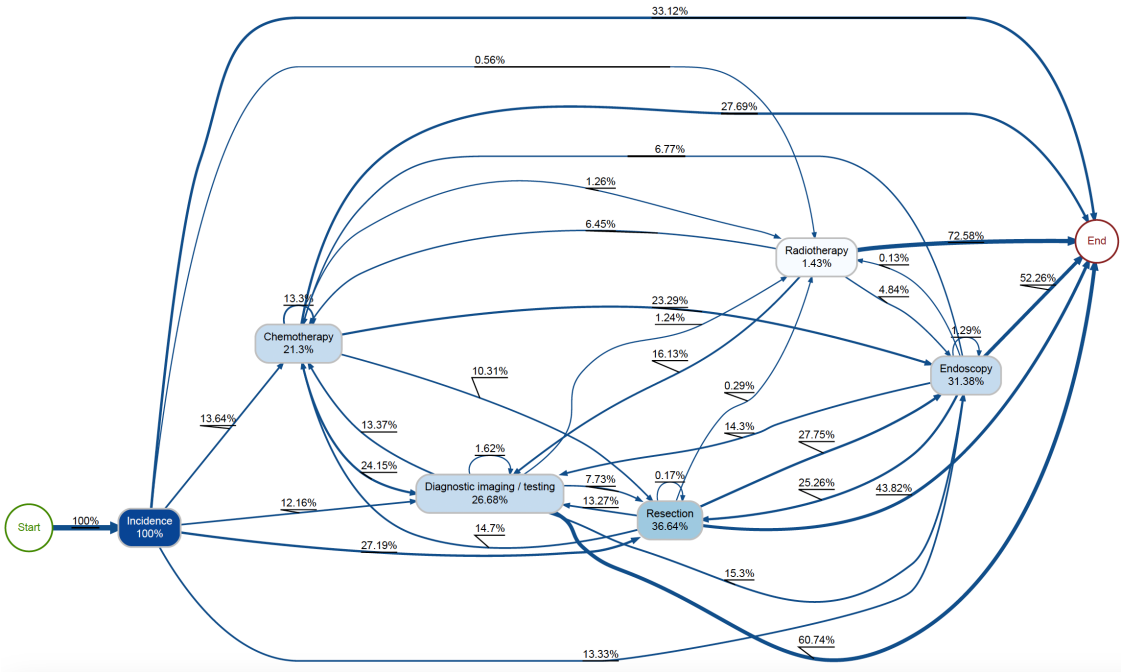


Figure 4.7: Directly-follows graphs of colon cancer treatment journeys at a regional NHS Trust (Dataset A) generated using activities selected by reference Codelist E

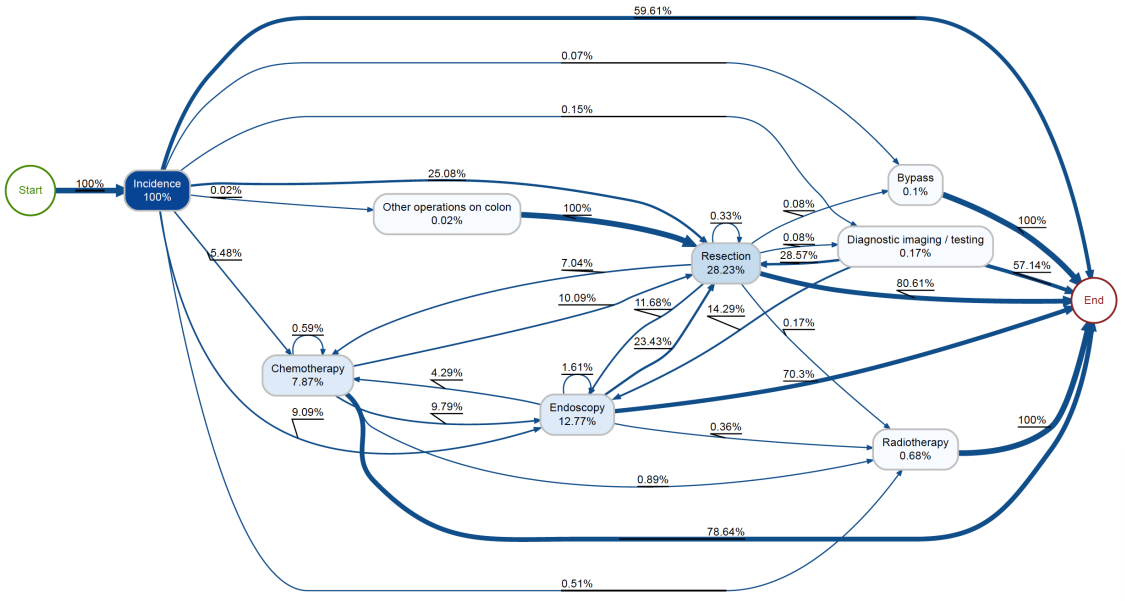


Figure 4.8: Directly-follows graphs of colon cancer treatment journeys at a regional NHS Trust (Dataset A) generated using activities selected by Query 1's SNOMED code-list

4.5.2.2 Lung Cancer

In lung cancer, there was a higher level of agreement between the two process models shown in Figures 4.9 and 4.10. In particular, a very similar rate of lobectomies were recorded by both models (10%). Chemotherapy and radiotherapy rates were also similar. An interesting difference is that the Query 1 code-list returned more codes related to sleeve/wedge resections as the rate of patient's with that procedure was 7.2% versus 2.5% by reference Codelist D. The smaller scale of discrepancies between these two models could be due to the smaller number of possible codes for lung cancer treatment as compared to colon cancer.

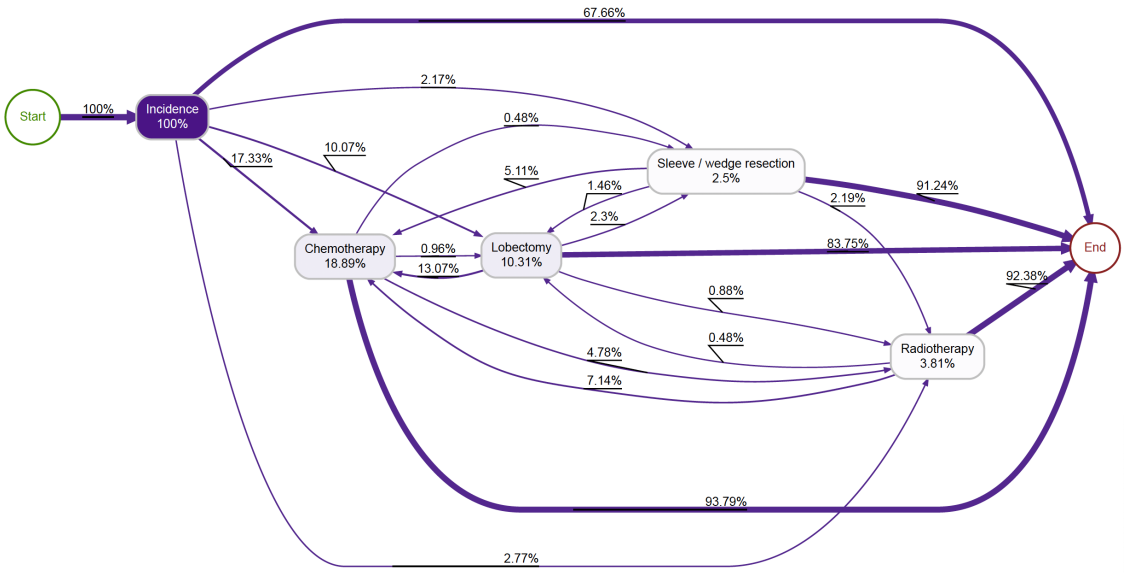


Figure 4.9: Directly-follows graphs of lung cancer treatment journeys at a regional NHS Trust (Dataset A) generated using activities selected by reference Codelist D.

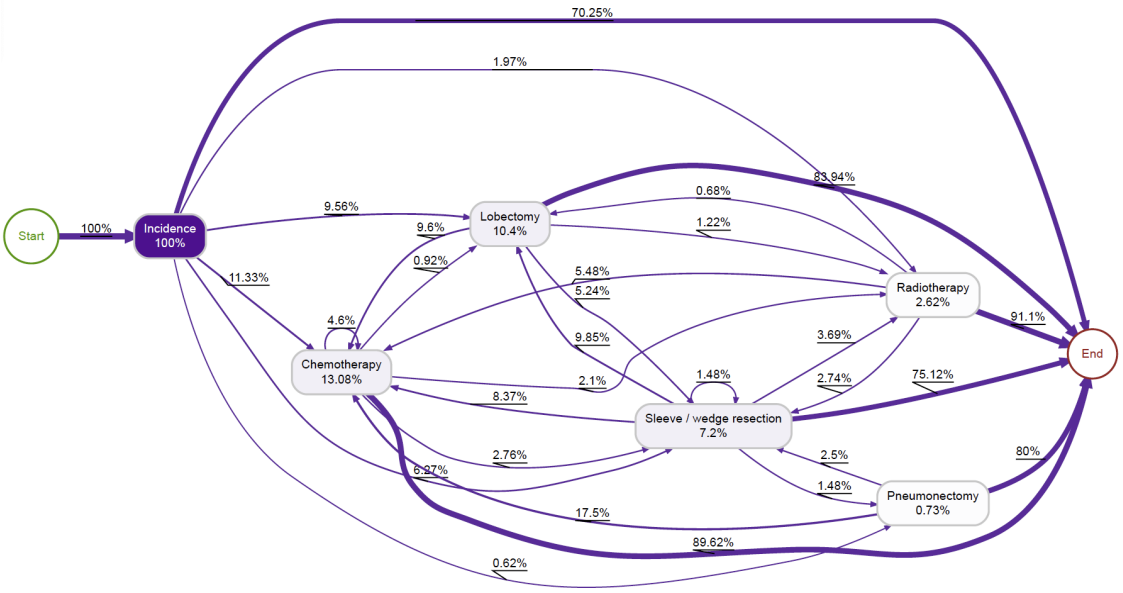


Figure 4.10: Directly-follows graphs of lung cancer treatment journeys at a regional NHS Trust (Dataset A) generated using activities selected by Query 1's SNOMED code-list

4.5.3 Enriching Code-lists with Statistics

The number of ICD codes found in at least one patient in the colon cancer case cohort was 1,394 (11% of all possible ICD codes found in MIMIC). Of these, 4% of the terms were in the SNOMED Query 1 list. If we restrict the codes to having a case prevalence of at least 0.96, or a minimum of 10 patients having the code, coverage of the SNOMED Query 1 code-list goes up to 13% of all possible codes.

The number of ICD codes found in at least one control cohort patient was 12,541 (99% of all ICD codes in MIMIC). Of these, 1.3% of the terms were in the SNOMED Query 1 list. If we restrict the codes to having a case prevalence of at least 0.96, or a minimum of 10 patients having the code, code-list coverage only increases up to 2%.

This difference in code-list coverage between cohorts shows that the SNOMED ECL Query lists are tuned to the queried disease, as precision is higher in the case cohort, and therefore will discern the pathway for the disease site of interest and should not pick up activities meant for other disease pathways. However, the clinical coding terminology being used will change the kinds of activities that are available to be mined, as will local clinical code practices. Therefore, it may be appropriate to supplement ontologically derived code-lists with codes from the data being analysed based on statistical methods to ensure complete coverage of possible events to be mined for a patient's journey.

Each ICD code was counted in the diagnosed (case) and non-diagnosed (control) population, and the rate of occurrence was compared by calculating an odds ratio of appearing in the case cohort and through statistical testing. The procedure codes with the strongest association (statistically ($p < 0.05$) more likely to appear in the case cohort, with an odds ratio of > 10) to a colon cancer diagnosis and are not in the SNOMED Query 1 code-list are presented in Table 4.3.

Most of the "missing codes" were related to the large intestine. The SNOMED queries did not return these as they are not tagged with a semantic relation marking the colon as a "a descendant of" the procedure site. The exception here is *endoscopic insertion of colonic stent(s)*, which does not have a recorded mapping to a particular SNOMED code in the OMOP data model.

This method of searching for procedural codes statistically correlated with a particular patient cohort can enhance ontology derived, or even domain expert derived, code-lists by identify missing codes that are correlated with the disease of interest but not included in the initial code-list due to, for example, local deviations of the patient journey or differences in clinical coding practices.

ICD concept		Prevalence (%)		Odds ratio	p
Code	Description	Control	Case		
0DTF4ZZ	Resection of Right Large Intestine, Percutaneous Endoscopic Approach	0.05	6.69	137.57	<0.05
0DTG4ZZ	Resection of Left Large Intestine, Percutaneous Endoscopic Approach	0.01	1.15	82.53	<0.05
4593	Other small-to-large intestinal anastomosis	0.15	7.65	55.54	<0.05
0DTF0ZZ	Resection of Right Large Intestine, Open Approach	0.07	3.25	51.38	<0.05
4594	Large-to-large intestinal anastomosis	0.08	2.87	34.97	<0.05
4686	Endoscopic insertion of colonic stent(s)	0.05	1.53	31.29	<0.05
0DJD4ZZ	Inspection of Lower Intestinal Tract, Percutaneous Endoscopic Approach	0.06	1.15	20.93	<0.05
4579	Other and unspecified partial excision of large intestine	0.06	1.05	16.46	<0.05
0D1B0Z4	Bypass Ileum to Cutaneous, Open Approach	0.18	2.58	15.08	<0.05

Table 4.3: Procedure codes most strongly associated to patients with a colon cancer diagnosis in the MIMIC dataset that did not also appear in the SNOMED Query 1 code-list.

4.6 Discussion

This chapter proposes utilising the domain knowledge stored in the SNOMED clinical terminology ontology to automatically generate lists of clinical codes, and therefore activities, for event-log generation and process discovery by querying SNOMED's Expression Constraint Language. The queries extract relevant codes by searching over the relationships between diagnosis and procedure codes annotated in the SNOMED ontology.

Study 1 and Study 2 establish that the results of a logical query over SNOMED's domain knowledge is capable of approximating a human-prepared procedural code-list with some accuracy. There exists a clear trade-off between precision and recall in designing the query: either a relatively complete code-list is returned with a large number of additional codes, or an incomplete code-list is returned in which every code has a high chance of being relevant. Additional procedure codes seen during the studies included those related to metastatic sites or un-related procedures where the queried body structure was used.

Additionally, performance varies depending on the type of event: surgeries are often easy to identify owing to their clear semantic annotations in the SNOMED ontology, whereas relevant but non-disease-specific activities such as chemotherapy, diagnostic tests, and imaging need to be explicitly specified. Query constraints, therefore, should be guided by the research question and purpose.

A benefit of the query process is that specifying a set of requirements for desired activities in terms of their relationship to a particular target disease keeps the process mining method focused on a concrete health issue as suggested by the epidemiology-informed process mining method (Chapter 3). It also generates a starting point for domain expert-led discussions, hopefully decreasing time spent preprocessing data to generate event-logs for patient journey mapping.

Lastly, this approach is generalisable to any EHR dataset coded in any clinical terminology– or at least any terminology with a mapping to SNOMED. Thanks to increased emphasis on interoperability, mappings between terminologies are well documented and available online, including tools created by NHS Digital and

OHDSI's standardised vocabularies [92], [107]. Translation quality from SNOMED into more localised clinical terminologies is dependent on both the completeness of the original terminology and the quality of mappings available. This issue was a limitation experienced in both studies, which used different mapping tools, each with its own discovered limitations (e.g. one to many mappings and loss of specific codes).

Abstracting events into meaningful and useful categories remains a key challenge in process mining and healthcare research. At present, the structure of the key ontologies used in healthcare data do not easily support activity abstraction methods; human curation is still required. Effective abstraction for research also depends on the research question: modelling an overarching treatment process requires a different abstraction level versus determining process differences within a particular aspect of a treatment pathway. For example, examining diagnostic pathways will require mapping individual diagnostic strategies whereas mapping an overview of a pathway may involve abstracting these strategies into the higher-order activity "diagnostic tests."

This chapter's proposed activity selection approach is far from a replacement for discussion and input from domain experts. In trying to analyse and then use these code-lists for process mining, the importance of having a deep understanding of the subject area, the research question, and the nature of the EHR data being used was highlighted. Without an understanding of the subject area, there was no way to quality check the code-lists themselves, and without an understanding of the provenience of the EHR data, there was no way to understand why some codes may or may not be seen in the data itself.

Data preprocessing stages are a key part of the process mining framework. Modelling assumptions and decisions, such as which events in a raw health record are included and which are discarded, inevitably impact the outcome of any analysis. Current practice has significant flaws: filtering an event-log based on the frequency of events risks losing useful information, and hand-curating code-lists is, whilst reliable, exceptionally time-consuming and leads to variable results in studies when lists are not shared.

Currently, no single basic approach, whether reliant on ontologies, on a conversation with domain experts, or on pure statistics, can fully capture the complex nature of healthcare processes; a well-rounded approach encompassing all three perspectives is required to deeply explore health processes and answer questions around modern healthcare delivery.

Future work should focus on defining more clearly an approach that combines all three of these methods for effective event-log generation from EHR data for process mining for healthcare. There should also be increased pressure for researchers, both in process mining and in epidemiology, to share their code-lists so that others may use them, ensuring reproducibility and comparability of results.

"There is an opportunity to develop and implement an ambitious dedicated plan for cancer so England can achieve world-class cancer services, and world-leading bowel cancer survival." -Bowel Cancer UK [64]

5

Understanding Stage IV Rectal Cancer Care Pathways

Contents

5.1 Introduction 82
5.2 Analysing Rectal Cancer 83
5.3 Methods 85
5.3.1 Defining the Patient Population and Subcohorts 85
5.3.2 Identifying Events of Interest 85
5.3.3 Event Log Generation and Patient Journey Classification 88
5.3.4 Analysis 88
5.4 Results 89
5.4.1 Patient Demographics and Journey Breakdowns 89
5.4.2 Patient Journeys 91
5.4.3 Factors Affecting Treatment Completion 97
5.4.4 Treatment Pathways by Cancer Alliance and MDT . . . 98
5.4.5 Effects of Complete Pathway Order 103
5.5 Discussion 112

5.1 Introduction

This chapter modifies epidemiology-informed process mining to present an application of process-informed epidemiology. In Chapter 3, the approach focused on embedding epidemiology methods into process mining analysis to ensure the

relevancy of results. Here, epidemiology analysis methods are used to analyse processes discovered using process mining pre-processing techniques, thereby expanding epidemiological analysis.

The epidemiology-informed process mining method, outlined in 3, is still relevant in process-informed epidemiology: the only modification is the methods used to compare results and investigate deviations. In this chapter, logistic regression, funnel plots, and Kaplan Meier survival analysis are used to compare patient journeys between patient groups and NHS Trusts.

The research questions to be answered in this chapter are:

- What are the treatment pathways for stage IV rectal cancer patients?
- How many patients have received a complete treatment pathway?
- Are there any factors associated with a patient's likelihood of receiving a complete treatment pathway?
- Do treatment pathways vary significantly between MDTs and cancer alliances?
- Does treatment activity order affect survival outcomes in patients who had a complete treatment journey?

Through answering these questions, this chapter will show the importance of expanding clinical research methods to consider whole processes when investigating care delivery.

5.2 Analysing Rectal Cancer

Stage IV cancer is characterised by the spreading, or metastasis, of the cancer from the original (primary) site to other organs or areas of the body.Metastasis occurs when parts of the primary cancer break away and spread to other organs through the bloodstream, causing them to grow and form new tumours. Rectal cancer metastasis sites are most frequently the liver (25%), lung (10-15%), and peritoneum (4-13%) [108], [109].

Prognosis of cancer is closely linked to stage at diagnosis, with the worst outcomes being observed in those with stage IV disease [110]. Unfortunately, up to 20% of rectal cancer patients present at diagnosis with such metastatic disease and, given the poor prognosis, there is an urgent need to optimise their care [111].

Treatments for stage IV rectal cancer need to be multi-disciplinary, requiring surgery to the primary and metastatic tumours, systemic anti-cancer therapy, and often radiotherapy [111], [112]. Delivery and decision-making of which surgeries and oncological treatments a patient is eligible for can involve several cancer site-specific MDTs and referral pathways for treatment at many hospitals.

The sequence of these treatments is insufficiently researched, leading to no consensus on what constitutes an optimal care pathway and care decisions devolved to individual clinicians or MDTs, increasing the chances of differences in treatment patterns between hospitals and concerns of inequalities in both care and outcome [113], [114].

Observational studies have identified unwarranted variation in single points on the care pathway, such as colorectal resectional surgery [17], adjuvant chemotherapy [80], radiotherapy [82], pulmonary metastasectomy [109], and liver resection [115]. These variations are seen between treating hospitals and between patient groups. For example, it is well documented that patients living in more deprived areas receive less treatment than patients living in more affluent areas [2], [116], [117]

Due to the complexity of treatment for stage IV disease, these studies may only reflect a small fraction of care differences. This chapter aims, therefore, to combine process modelling and epidemiological techniques to quantify and explore variation across the full pathway of stage IV rectal cancer, and relate this variation to outcomes, across the English NHS. Current national practice is quantified, patient determinants of receiving a complete treatment pathway are investigated, variation between practices is explored, and finally pathway differences are linked to survival.

5.3 Methods

To research stage IV rectal cancer patient pathways, the epidemiology-informed process mining approach outlined in Section 3.2 was utilised.

5.3.1 Defining the Patient Population and Subcohorts

The datasets used were the national cancer registry, HES, SACT, and RTDS, which were all linked at the patient level and accessed within the CORECT-R secure data environment [72].

Information on patients' date of diagnosis, age, sex, socioeconomic deprivation (derived from the income domain of the Index of Multiple Deprivation (2015)), and date of death or date of censorship, (the last time the patient was checked in the spine by NCRAS) were extracted from the cancer registry dataset. Every individual in the cohort was allocated to a primary MDT of management based on the hospital in which they underwent their primary resection or, if there is was no primary resection, the hospital they attended closest to diagnosis; each MDT is assigned to a Cancer Alliance.

The patient population of study consists of 12,827 patients diagnosed with stage IV rectal cancer in England between 1 January 2012 and 31 December 2019. Patient sub-cohorts are defined based on demographics (sex, age-group, IMD, and Charlson comorbidity score) and treating MDT and Cancer Alliance.

5.3.2 Identifying Events of Interest

All episodes of secondary care were identified using the linked HES, SACT, and RTDS datasets. Activities such as creation of stoma, insertion of stent, surgical and ablative treatments to both primary and metastatic sites were derived from HES; SACT regimens were derived from SACT; and radiotherapy regimens were derived from RTDS. For every individual in the cohort a patient journey was then created by following their interventions over time.

To select and filter for the correct surgical, endoscopic or ablative interventions in HES, and then relabel and group them into activities of interest, a data

dictionary was built that maps OPCS codes to treatment names. Labels of differing granularity were derived in the data dictionary, including grouping surgical procedures into primary site directed treatments and metastatic site directed treatments. The surgical data dictionary was created in collaboration with domain experts (oncologists, surgeons, cancer epidemiologists) and was guided by the CORECT-R data coding dictionary [72] and published online to Github [85].

All surgical events in the data dictionary that occurred one month before a patient's incidence date and any time after were selected from the inpatient and outpatient HES data-tables. For each surgical event, the admission method for the patient, the hospital name, and the surgery event date were also selected.

Rectal radiotherapy was grouped using an established algorithm into short-course radiotherapy with immediate surgery, short-course radiotherapy with delayed surgery, long-course chemo-radiotherapy, post-operative radiotherapy or another regimen [82]. To select for radiotherapy events, the treatment start date of a unique radiotherapy episode for each patient was selected along with the treating hospital code.

To select for SACT events, the first treatment date of each unique regimen was selected for each patient, along with the hospital code. The selected regimens were then filtered for colorectal oncological drugs and categorised into five groups, shown in Table 5.1.

The filtered datasets were then merged and arranged by date to generate the finalised event-log. All patient journeys start with date of diagnosis and end with either the date of death or date of censorship (labeled as "alive"). The high-level pathway treatment activities, with their sub-categories, are defined as:

- **Primary Management:** consisting of the surgical activities stoma, stent, minor resection, and bypass
- **Primary Resection:** consisting of the surgical activities anterior resection, Hartmanns resection, exenteration, abdominoperineal resection, and other colectomy

- **Metastasis Treatment:** consisting of the surgical activities ablation or resection to a liver, lung, adrenal, peritoneum, and/or CNS site
- **Chemotherapy Regimen Start:** the start date of a new chemotherapy regimen as defined by the SACT dataset. The regimens are categorised into five regimens or grouped into the singular event "SACT."
- **Radiotherapy Regimen Start:** the start date of a new radiotherapy regimen as defined by the RTDS dataset. Radiotherapy episodes are classified into five categories using the algorithm defined in Morris, Finan, Spencer, *et al.* [82], which are then grouped either into treatment to the primary tumour (definitive and neo-adjuvant) or treatment to the metastatic tumour (definitive and SBAR).

Drug Regimen Listed in SACT	**Regimen Analysis Group**
Capecitabine monotherapy Capecitabine monotherapy + anti-VEGF Fluorouracil monotherapy Fluorouracil monotherapy + anti-EGFR Fluorouracil monotherapy + anti-EGFR	5FU/Cape +/- biological
CAPIRI CAPIRI + anti-EGFR CAPIRI + anti-VEGF CAPOX CAPOX + anti-EGFR FOLFIRI FOLFIRI + anti-EGFR FOLFIRI + anti-VEGF FOLFOX FOLFOX + anti-EGFR	5FU/cape based Doublet +/- biological
FOLFOXIRI FOLFOXIRI + anti-VEGF	Triplet +/- biological
Immunotherapy Other monotherapy Other monotherapy + anti-EGFR Other combination therapy Anti-EGFR Anti-VEGF	Other

Table 5.1: Systemic anti-cancer drug treatment regimen analysis groups.

5.3.3 Event Log Generation and Patient Journey Classification

The research questions focus on curative treatments. Therefore, for patient journey analysis, the primary management activities were dropped as their focus is the management of symptoms versus removing the rectal cancer. Radiotherapy events, based on site, were further grouped with primary resection events (to form primary treatment) and metastatic treatment events (to form metastatic treatment), based on the radiotherapy's intended treatment site, as the clinical goal is the same.

287 unique journeys at the highest abstraction level were mined. These were then grouped into 16 distinct, ordered journeys and then 8 distinct, un-ordered pathway groups, shown in Table 5.3. A distinct, ordered journey is generated by taking the first instance of an activity class in the pathway while maintaining the order of events. A distinct, un-ordered journey takes the first instance of an activity class but does not maintain the order of the events. This means multiple SACT regimens or multiple surgeries are not account for. The choice for abstracting the pathways to a distinct, un-ordered journey is the focus of this investigation is on which activities a patient receives versus understanding the exact treatment pathway.

The treatment pathways were further grouped into three treatment types (Table 5.2): those who had no radical treatment were defined as having an incomplete pathway, those who had radical treatment to either their primary tumour or a metastatic site with or without SACT were deemed as having a partial pathway, and those who had radical treatment to both the primary tumour and metastatic site, again with or without SACT, were deemed as having a complete pathway.

5.3.4 Analysis

Rates of use of defunctioning stomas, insertion of stents, rectal radiotherapy, chemotherapy, and surgery for both primary and metastatic disease, and proportion of patients in the incomplete, partial and complete pathway groups were quantified both in relation to patient characteristics and across the NHS at both MDT and

Cancer Alliance level. Heat-maps were created to show the proportion of patients who followed each pathway by MDT and Cancer Alliance.

To illustrate the full patient journey, dot plots were created, with time on the x-axis marked in weeks from the date of diagnosis and the patient ID on the y-axis; different-coloured dots represent each clinical event on the pathway.

Multi-level logistic regression was used to determine the factors associated with completion of the treatment pathway. Models were constructed with patients clustered within MDTs, which were further clustered within Cancer Alliances [17]. Explanatory variables included age at diagnosis, sex, IMD quintile, year of diagnosis and Charlson comorbidity score. Funnel plots were constructed to show the variation in rates of complete pathways across MDTs [58].

To link treatment order with outcomes unadjusted survival was calculated from date of diagnosis to date of death or when censored. Survival probability by complete treatment pathway was assessed visually using the Kaplan-Meier method and using log rank tests for determining statistical significance [59]. Given that survival will inevitably be linked to the length of the treatment pathway, known as immortal time bias, a landmark survival analysis was also conducted using last date of treatment in the pathway as the entry point [118].

To visually assess adjusted survival differences between pathways, cox-proportional hazard regression models were created, calculating survival from diagnosis and last date of treatment. The models were adjusted for patient age group, sex, Charlson score, IMD, and year of diagnosis. The treatment pathway was included as a stratifying factor to ensure each group has its own baseline hazard.

5.4 Results

5.4.1 Patient Demographics and Journey Breakdowns

Table 5.4 presents descriptions of the patient population including demographics and the number of metastatic sites identified for each patient. The population is majority male (65%) and white (90%), with a roughly equal split between indexes

of deprivation and age-group. Metastatic sites could only be identified for the 22% of patients who received a metastatic directed treatment.

Tables 5.2 and 5.3 show the breakdowns of patient numbers by treatment type and treatment pathway. The largest group of patients had a partial treatment pathway type (46%); 39% had an incomplete treatment pathway, and only 14% of patients had a complete treatment pathway. The most frequent treatment pathways were first, only receiving at least one regimen of SACT (21%), second, receiving at least one regimen of SACT with ablation/surgery to a metastatic site and resection and/or radiotherapy to the primary site (12%), and third, receiving at least one regimen of SACT and resection and/or radiotherapy to the primary site (11%).

Please note that having an "incomplete" or "partial" pathway does not necessarily indicate suboptimal or poor-quality care. In many instances, the pathway may reflect the most clinically appropriate course of action for the patient. These discovered pathways are based on the care given, not the care intended to be given, and therefore they can be a reflection of side effects for the patient or even death. In future, having data on intended care plans and clinical decision making will provide the ability to analyse if care delivered is sub-optimal. This chapter focuses however on making use of the data that is available and studying treatment patterns and variation of care as it is recorded in electronic health records currently.

The patient population demographics and patient journey breakdowns are also analysed and grouped by cancer alliance, presented in Appendix Table B.

Outcome	Treatment Type	Number of Patients (N)	Proportion (%)
0	Incomplete Pathway	5,052	39
0	Partial Pathway	5,927	46
1	Complete Pathway	1,848	14

Table 5.2: The number and proportion of patients who received either an incomplete, partial, or complete treatment pathway.

Pathway Activities	Treatment Type	N	%
Diagnosis Only	Incomplete Pathway	5,052	39.4
Primary Treatment	Partial Pathway	783	6.1
SACT	Partial Pathway	2,670	20.8
Metastasis Treatment	Partial Pathway	487	3.8
SACT and Primary,	PartialPathway	1,405	11.0
SACT and Metastasis	Partial Pathway	582	4.5
Primary and Metastasis	Complete Pathway	361	2.8
SACT, Primary, and Metastasis	Complete Pathway	1,487	11.6

Table 5.3: The number and proportion of patients in each treatment pathway group.

5.4.2 Patient Journeys

Dotted charts can be used to visually see how pathways differ between patients and patient groups by showing general patterns in care. Figures 5.1 show every patient's journey at two different levels of event abstraction. While some patterns can be seen– SACT is administered in the first 5 - 10 weeks from diagnosis, primary management events such as stomas happen first, and palliative radiotherapy occurs at the end of life– the journeys are highly variable between patients. Consistently though, high volumes of treatments are occurring till end of life and throughout the patient's journey.

Figure 5.2a shows the first year of treatments journeys for the youngest patients (aged $\leq$60 at diagnosis) while Figure 5.2b shows the first year of treatments journeys for the oldest patients (>80 at diagnosis). The graphs have very different shapes due to the difference in survival time and the volume of treatments received between these patient groups. For those under 60, we can see more kinds and rounds of treatment received compared to those over 80.

Given how complex and varied the treatment processes are for stage IV rectal cancer patients in England, rates of specific treatments were analysed before investigating process variation as a whole. The journeys and events analysed include those with non-curative intent, such as primary management activities and palliative radiotherapy.

Characteristic	Number of Patients	Proportion (%)
Sex		
Male	8,386	65
Female	4,441	35
Age Group		
≤60	3,248	27
61-70	3,344	26
71-80	3,554	28
>80	2,501	19
Ethnicity		
White	11,571	90
Other	665	5.2
Unknown	591	4.6
IMD (2015)		
1- Most affluent	2,489	19
2	2,736	21
3	2,678	21
4	2,545	20
5- Least affluent	2,379	19
Charlson Comorbidity Score		
0	10,085	78.7
1	1,627	12.7
2	609	4.7
3	506	3.9
Treated Metastatic Site		
Liver	1,439	11.3
Lung	486	3.8
Oligo-sites	567	4.4
Other	374	2.9
Unknown	9,961	77.6
Number of Treated Metastatic Sites		
1	2,299	17.9
2	491	3.8
3	75	0.58
4	<5	0.21
Unknown	9,961	77.6

Table 5.4: Patient demographics for all 12,827 stage IV rectal cancer patients analysed.

5.4.2.1 Events Along the Journey

Surgical Activities: 58% of patients had at least one surgical event (primary management, primary resection, metastatic ablation/resection) in their pathway. Of those, 41% (24% of all patients) had a resectional event to the primary tumour and 38% (22% overall) had an ablative or resectional event to a metastatic site. Overall, 17% had a surgical event to both tumour sites. Anterior resection was the most frequent surgical activity (27% of those with a surgery event), followed by ablation or resection to a liver metastasis site (25% of those with a surgery event).

Rectal Cancer Management Activities - Stomas and Stents: Stoma formation was the most frequent first surgical treatment on a patient's pathway (35.8%), and overall, 36.8% of patients had a defunctioning stoma event. Of those patients, 46% had an additional resectional event for their rectal cancer, but 12% only received a stoma as their only rectal cancer surgical activity. 4% of patients had a colonic stent inserted, with it being the first rectal cancer surgical event in 74% of patients.

SACT Regimens: 48% of patients had at least one regimen of systemic anti-cancer therapy, with 43% of those having it as their only treatment, 34% having it before surgery or radiotherapy, and 22% having it after surgery or radiotherapy.

Table 5.5 shows for each SACT regimen the number and proportion of unique patients who have it as their first regimen, or are treated with that regimen at some point in their journey. Appendix C shows patient characteristics by first SACT regimen.

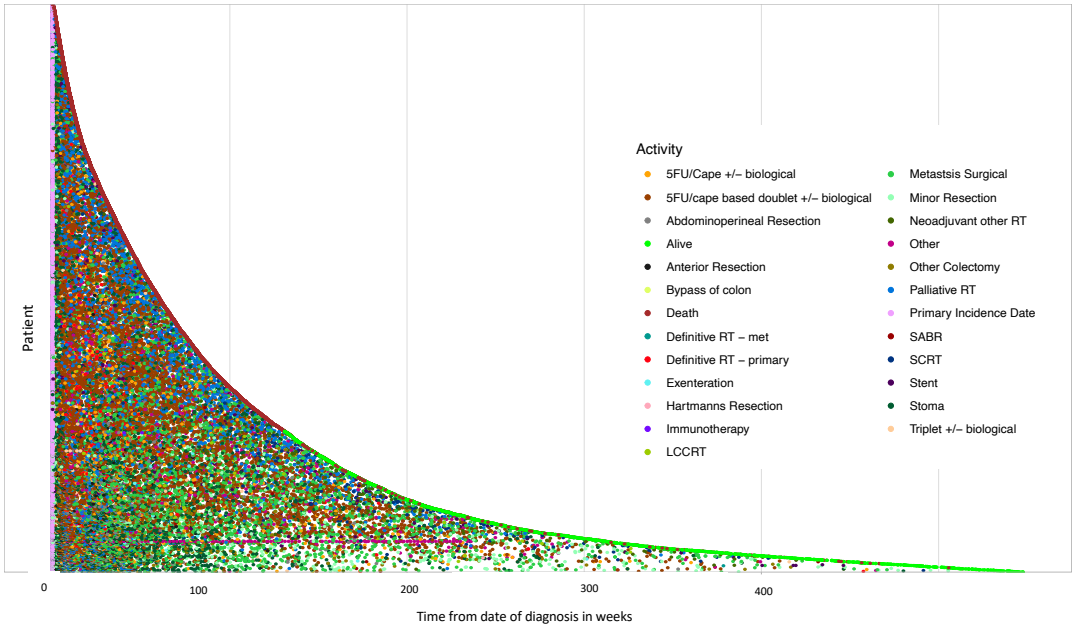
SACT Regimen	First Regimen		Regimen at any time	
	N	%	N	%
5FU/cape based doublet +/- biological	4,541	73.9	5,316	61.5
5FU/Cape +/- biological	1,167	19.0	2,013	23.3
Other	335	5.5	1,162	13.4
Triplet +/- biological	98	1.6	147	1.70
Immunotherapy	<5	0.05	7	0.08

Table 5.5: Number (n) and proportion (%) of patients with a specific SACT regimen either given as their first regimen, or given at any time in their treatment journey.

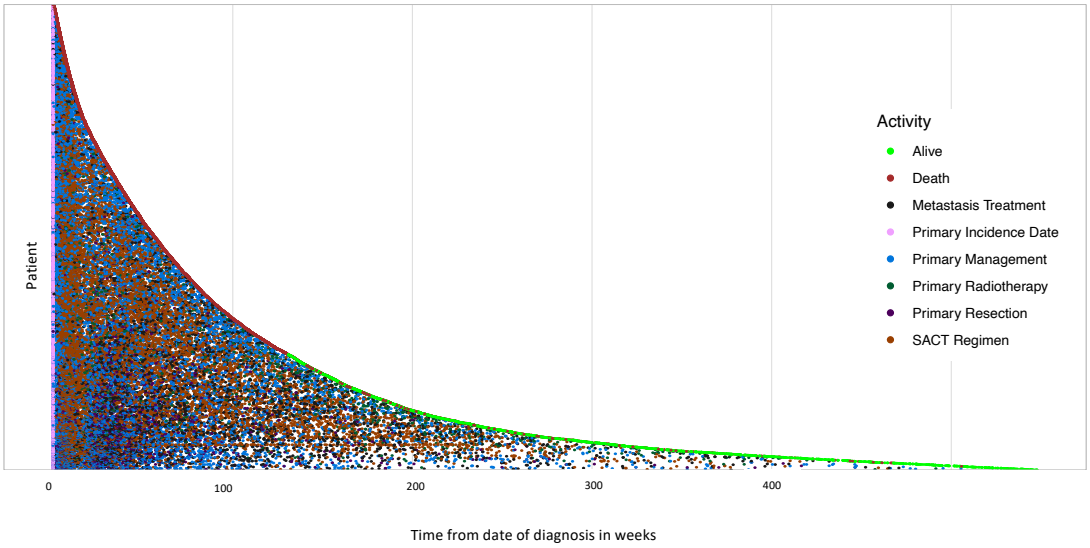
Radiotherapy: 43% of patients received some form of radiotherapy treatment, with the kinds of radiotherapy broken down into regimens in Table 5.6. The most frequent type of radiotherapy given was palliative, with 71% of patients receiving at least one round of palliative radiotherapy.

Radiotherapy Regimen	**Number of Patients** (N)	**Proportion of Patients (%)**
Palliative	4983	71.1
Definitive - rectal site	976	13.9
Neoadjuvant other	439	6.3
Short-course radiotherapy	367	5.2
Long-course chemoradiotherapy	128	1.8
SABR	110	1.6
Definitive - metastatic site	9	0.1

Table 5.6: The absolute number and proportion of patients with a specific radiotherapy regimen.

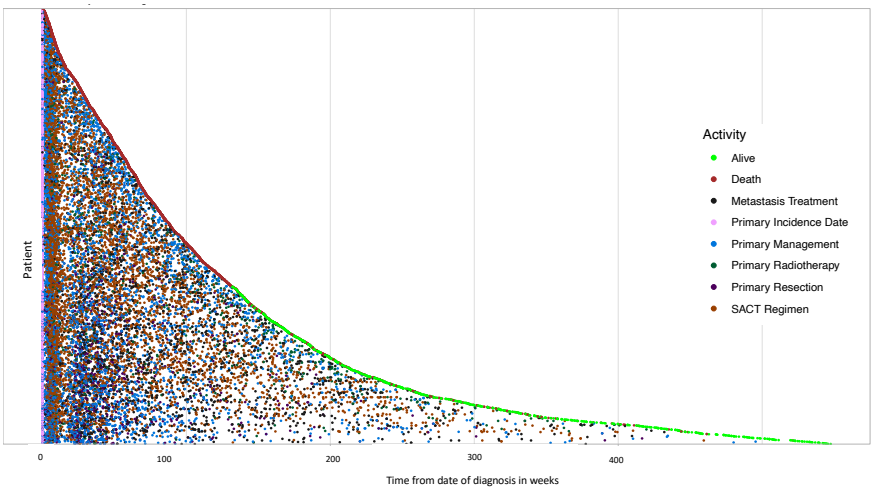


(a) Lowest abstraction level

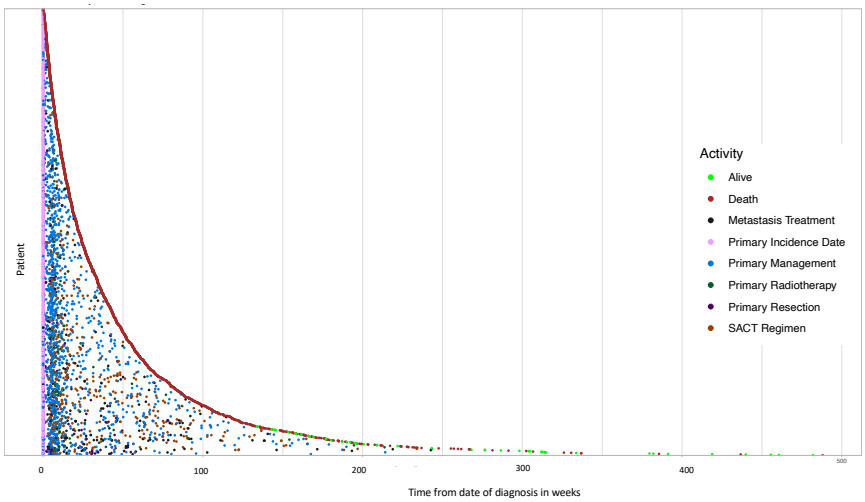


(b) Highest abstraction level

Figure 5.1: Dotted charts showing treatment events present in the event-log for each patient, at different activity abstraction levels. Each patient is marked along the y-axis; the x-axis represents time in weeks from date of diagnosis.



(a)



(b)

Figure 5.2: Dotted charts restricted to the first year after diagnosis and to patients aged 60 or less (a) and more than 80 (b) at diagnosis. Each patient is marked along the y-axis; the x-axis represents time in weeks from date of diagnosis.

5.4.3 Factors Affecting Treatment Completion

Table 5.7 contains results of the multi-level logistic regression to predict if a patient will have a complete treatment pathway. Incomplete treatment and partial treatment pathway types were grouped together to compare against complete treatment type pathways.

Women, those with increased age, those with greater comorbidity, and those living in more socioeconomically deprived areas were all significantly less likely to experience a complete treatment pathway. Year of diagnosis does not have a significant effect on whether a patient has a complete treatment pathway. These results are in-line with previous literature showing that those who are younger, richer, and male will receive more treatment than their counterparts.

Characteristic	**OR**	**95% CI**	**P value**
Sex			
Male	1.00		
Female	0.80	0.72 - 0.90	<0.01
Age Group			
≤60	1.00		
60-69	0.76	0.67 - 0.86	<0.01
70-79	0.42	0.37 - 0.48	<0.01
80	0.07	0.06 - 0.10	<0.01
IMD (2015)			
1- Most affluent	1.00		
2	0.87	0.75 - 1.00	0.06
3	0.83	0.71 - 0.96	0.02
4	0.66	0.56 - 0.77	<0.01
5- Least affluent	0.62	0.53 - 0.74	<0.01
Charlson comorbidity score			
0	1.00		
1	0.62	0.53 - 0.74	<0.01
2	0.26	0.18 - 0.38	<0.01
3	0.18	0.11 - 0.29	<0.01
Year of diagnosis	0.98	0.96 - 1.00	0.07

Table 5.7: Results of the multi-level logistic regression to predict if a patient will have a complete treatment pathway. Odds ratios show the increase or decrease of the odds of a patient having a complete treatment pathway compared to the reference group, whose odds are 1, or with one increase of unit if the characteristic is numerical.

5.4.4 Treatment Pathways by Cancer Alliance and MDT

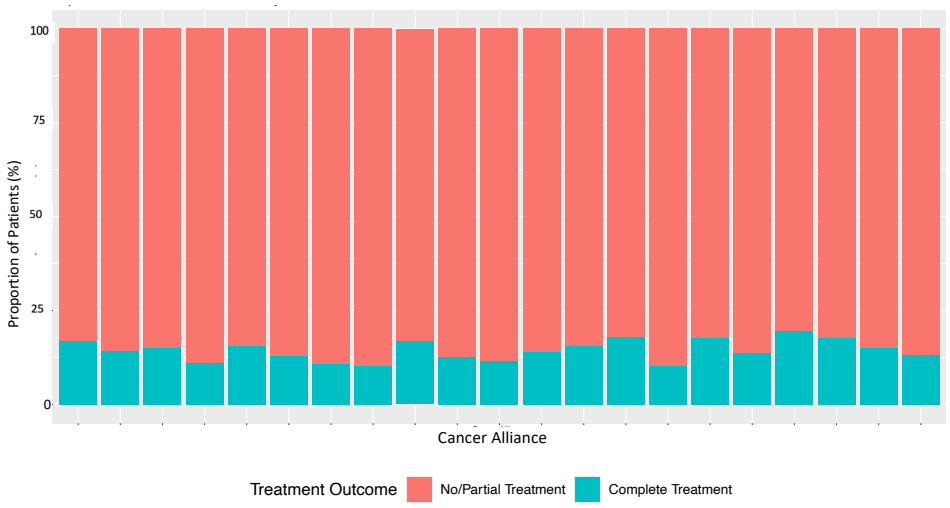
Figure 5.3 shows the proportion of patients with a complete or not complete pathway and the breakdown of patients between each treatment type by Cancer Alliance. Figure 5.4 shows these breakdowns by MDT.

The proportion of patients with incomplete treatment varied from 29% to 51%, with partial treatment from 39% to 55%, and with complete treatment from 10% to 20% between Cancer Alliances. The variation was even greater between MDTs where the proportion of patients with incomplete treatment varies from 7% to 64%, with partial treatment from 19% to 75%, and with complete treatment from 5% to 36%. NHS England's average for incomplete treatment is 39%, partial treatment is 46%, and complete treatment is 14%.

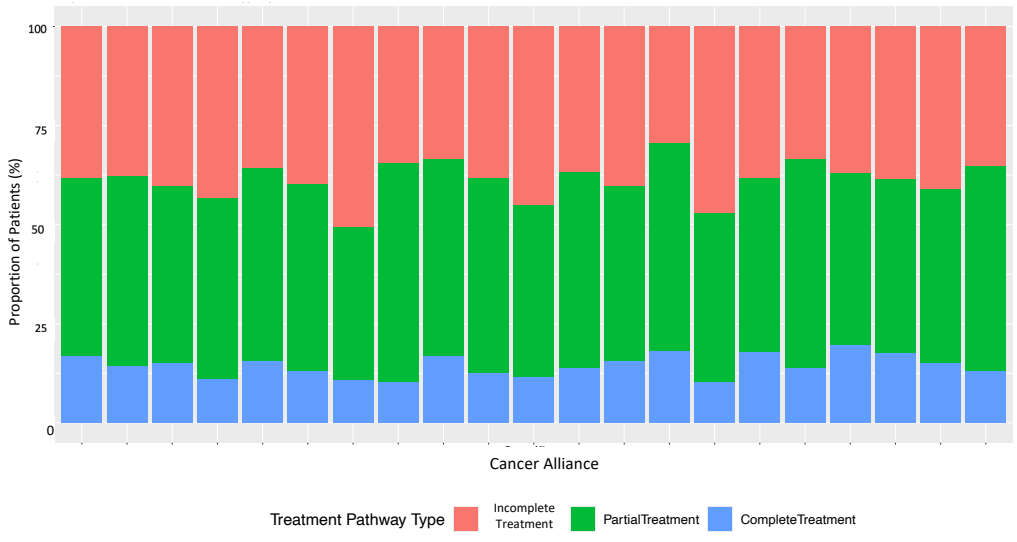
To verify if the difference in rates of treatment completion between MDTs is not due to chance, funnel plots were used to compare rates of treatment completion for each MDT while accounting for differences in patient populations. Figure 5.5a shows unadjusted funnel plots and Figure 5.5b shows the adjusted funnel plots.

The unadjusted funnel plot shows 3 outliers which are all large cancer referral centres. Risk-adjustment of the funnel plots brings all MDTs into the 95% confidence interval. This indicates that MDTs, when adjusting for their population case-mix, have patients with complete treatment pathways at similar rates.

Breaking down the treatment types into 8, un-ordered distinct treatment pathways reveals variation in the proportion of patients in each Cancer Alliance and MDT on them, shown by Figures 5.6a and 5.6b respectively. Variation between Cancer Alliances is not as stark as between MDT as a Cancer Alliance is the average of many MDTs. Some MDTs show clear preference for particular treatment pathways, with over 50% of their patients receiving a particular pathway, such as SACT and treatment to the primary site, or even the singular treatment pathway of receiving just SACT.

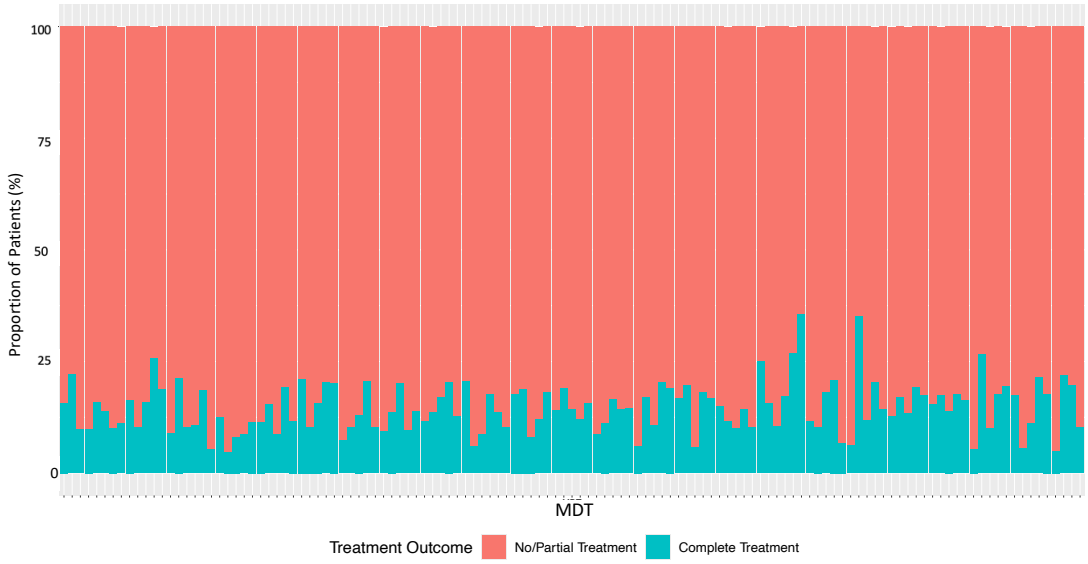


(a) Cancer alliance patient proportions by treatment completion

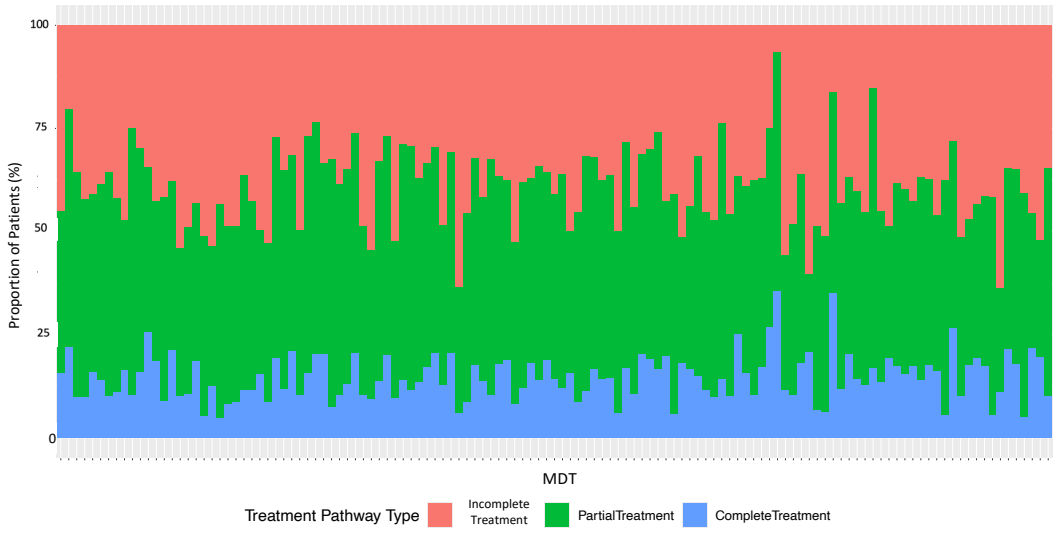


(b) Cancer alliance patient proportions by treatment completion

Figure 5.3: Proportion of patients in each Cancer Alliance by (a) treatment completion and (b) treatment pathway type.

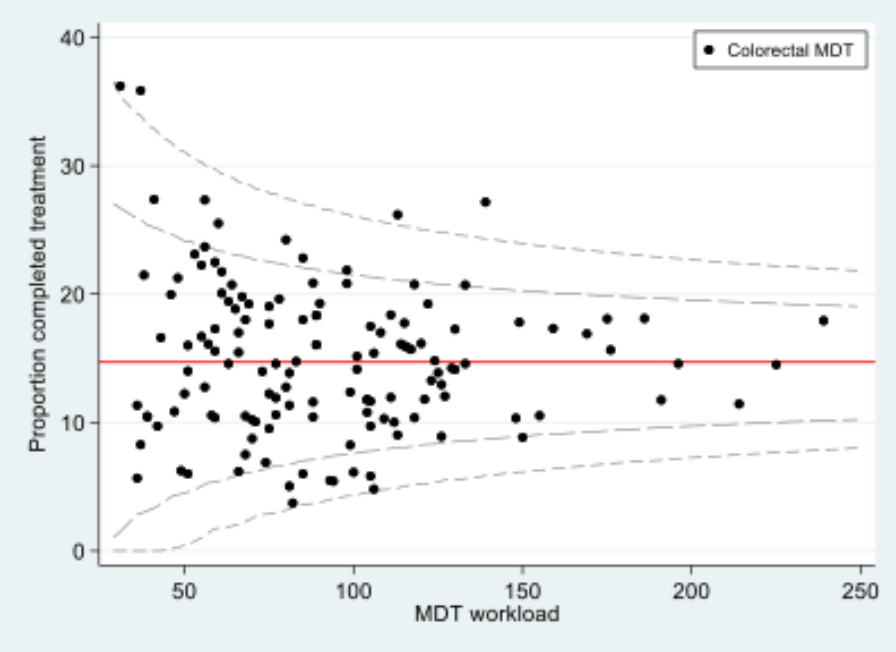


(a) MDT patient proportions by treatment completion

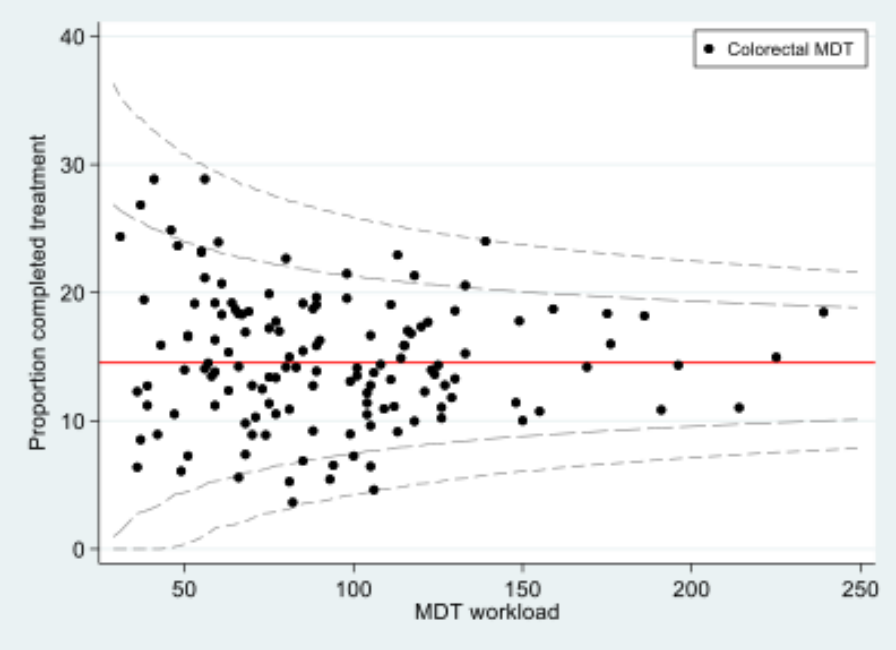


(b) MDT patient proportions by treatment pathway type

Figure 5.4: Proportion of patients in each MDT by (a) treatment completion and (b) treatment pathway type.

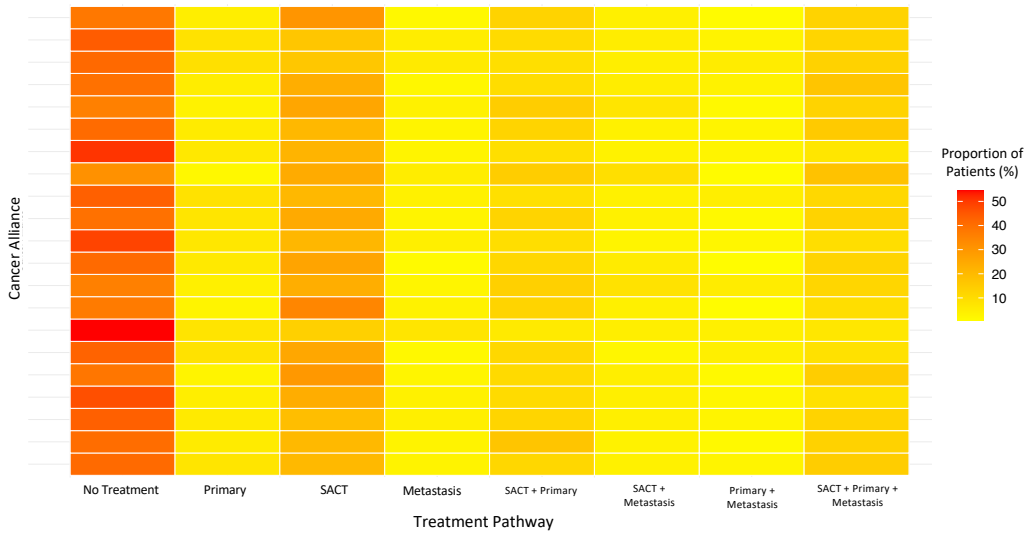


(a) Unadjusted Funnel Plot

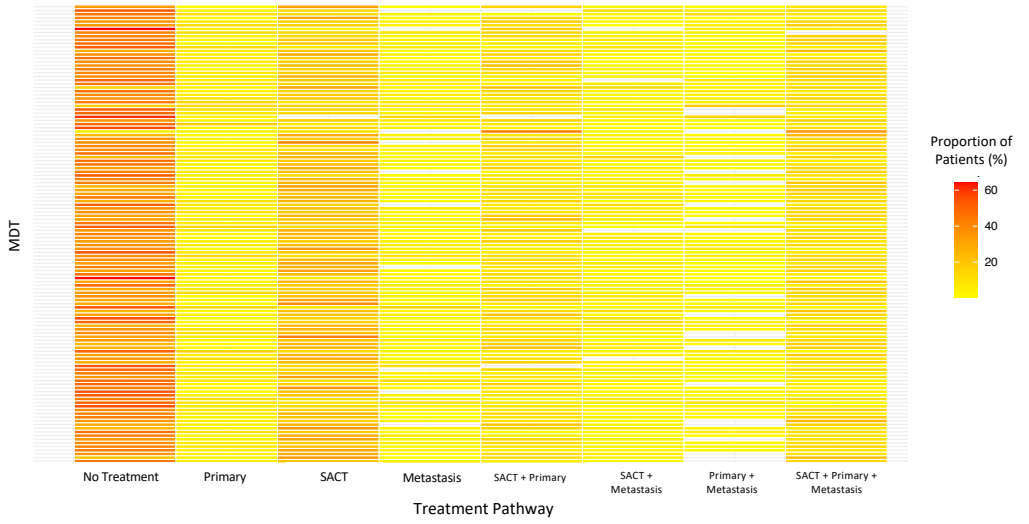


(b) Adjusted Funnel Plot

Figure 5.5: Funnel plots to compare treatment completion rates between MDTs.



(a) Treatment pathways by Cancer Alliance



(b) Treatment pathways by MDT

Figure 5.6: Heatmaps showing the proportion of patients in each treatment pathway by (a) Cancer Alliance and (b) MDT. The y-axis represents the treating resource and the x-axis represents the complete treatment pathway group. The proportion of patients at the resource on that pathway is coloured from yellow to orange to red as the proportion increases.

5.4.5 Effects of Complete Pathway Order

5.4.5.1 Variation by Cancer Alliance and MDT

To further show the benefits of a process-oriented approach, the order of complete treatment events and their relative effect on survival was investigated. There are 1,848 patients who received a complete treatment pathway. Within the two distinct, un-ordered complete treatment pathways, there are 8 distinct ordered pathway groups which a patient could be assigned. Table 5.8 shows the breakdown of patient numbers based on the order of events in their complete pathway. There is a large distribution of patients between the eight pathways.

Group	Treatment Activity Order	Number of Patients (N)	Proportion of Patients (%)
A	Metastasis, Primary	100	5.4
B	Metastasis, Primary, SACT	69	3.7
C	Metastasis, SACT, Primary	87	4.7
D	Primary, Metastasis	261	14
E	Primary, Metastasis, SACT	209	11
F	Primary, SACT, Metastasis	240	13
G	SACT, Metastasis, Primary	315	17
H	SACT, Primary, Metastasis	567	31

Table 5.8: The number and proportion of patients with a complete treatment pathway group broken down by treatment activity order.

Group H was the largest group with 31% of complete treatment patients receiving at least one regimen of SACT, followed by definitive treatment to the primary rectal tumour, followed by a definitive treatment to the metastasis site.

Group B was the smallest group with 3.7% of complete treatment patients receiving definitive treatment to the metastasis site, followed by definitive treatment to the primary rectal tumour, followed by systematic anti-cancer treatment.

Figure 5.7 shows heatmaps of the proportion of patients with each complete pathway group broken down by treating Cancer Alliance and MDT. The heatmaps show that there is large variation in the order of complete treatment pathway activities patients are receiving between Cancer Alliances. Variation is even greater

when examined by MDT, with many MDTs showing a preference for a specific treatment order. For example, between MDTs there can be 10% or 90% of their patients in Group H (SACT, Pri, Met). Some Cancer Alliances and MDTs don't even have patients on all of the complete treatment pathways.

5.4.5.2 Survival

To show how treatment process can be linked to outcomes such as survival, Kaplan-Meier curves were generated, with survival calculated from diagnosis and from date of last treatment, to account for the immortal time bias. Survival is calculated initially unadjusted and then adjusted for demographics (sex, age, Charlson score) and year of diagnosis.

Figures 5.8 and 5.9 show the KM curves, adjusted and unadjusted, for all 8 complete treatment pathway groups. Log-rank test of the unadjusted curves shows no statistical difference between curves (p=0.05) when survival is calculated from diagnosis, but there is a statistical difference when survival is calculated from last day of treatment (p=0.006). Adjustment for demographics increases survival slightly for most curves. There does, however, appear to be differences between similar pathways in all of the Figures. To investigate this further, the survival curves based on starting treatment activity were compared separately.

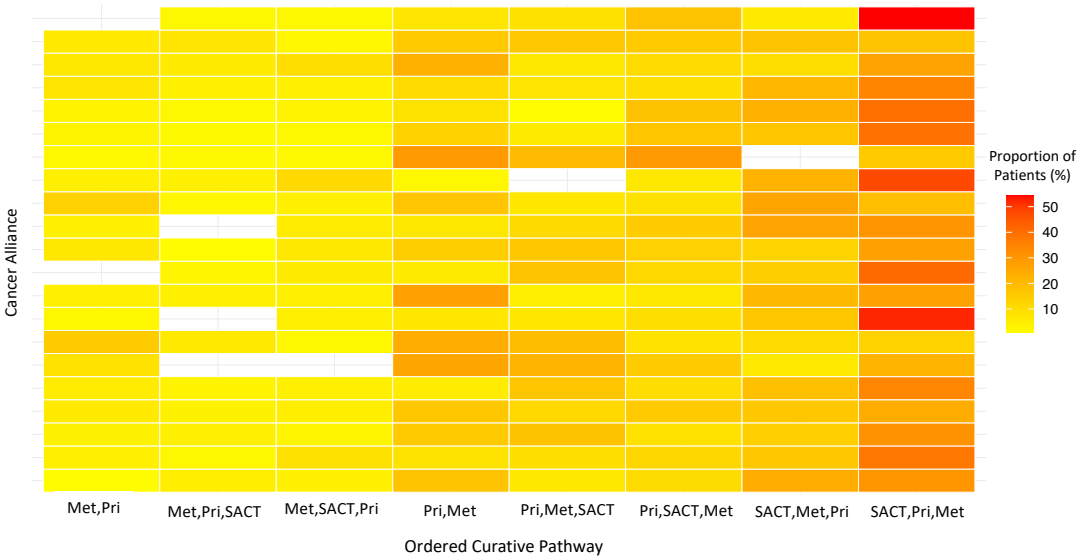
Unadjusted survival for patients on pathways starting with primary treatment (Groups D,E,F) is not statistically different when calculated from diagnosis (p=0.81) or from last date of treatment (p=0.31) (Appendix D).

There is also no significant difference in survival between pathways starting with metastasis treatment (Groups A,B,C) from diagnosis (p=0.82) or from last date of treatment (p=0.31) (Appendix E).

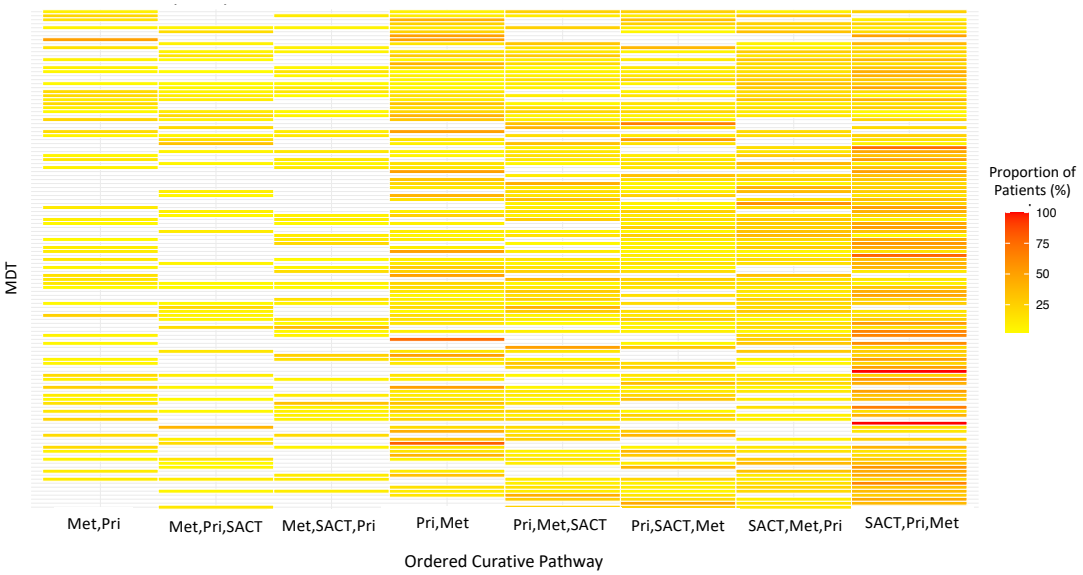
Groups G and H, pathways starting with SACT, have a significant difference in survival, both from diagnosis (p=0.02) and last date of treatment (p=0.003) (Figure 5.10). Survival is higher in patients who after SACT have metastatic site treatment directed treatment before treatment to their rectal cancer. The difference is largest when calculating survival from last date of treatment.

Patient characteristics for those in Group G and H are shown in Table 5.12. Overall characteristics for both groups are also presented for comparison.

The biggest difference between the groups is the number of metastatic sites and the metastatic site locations. Patients in Group G (SACT, Met,Pri) have metastasis in the liver (63%), or in multiple locations (31%), with almost no lung or other metastasis sites treated. Group H (SACT, Pri,Met), however, treats almost all patients with a lung metastasis in this cohort, and has a greater overall mix of metastatic sites treated.

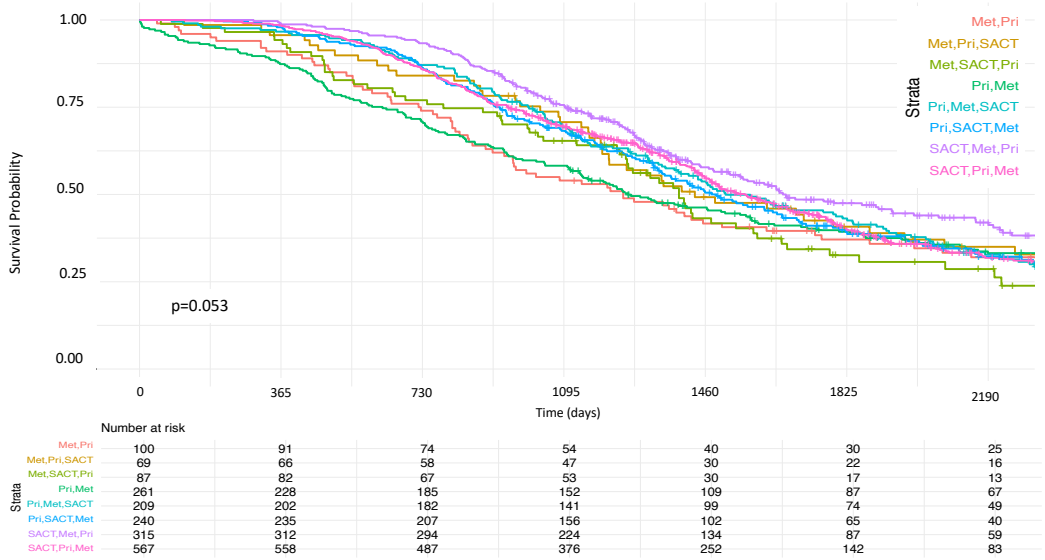


(a) Complete pathway order by Cancer Alliance

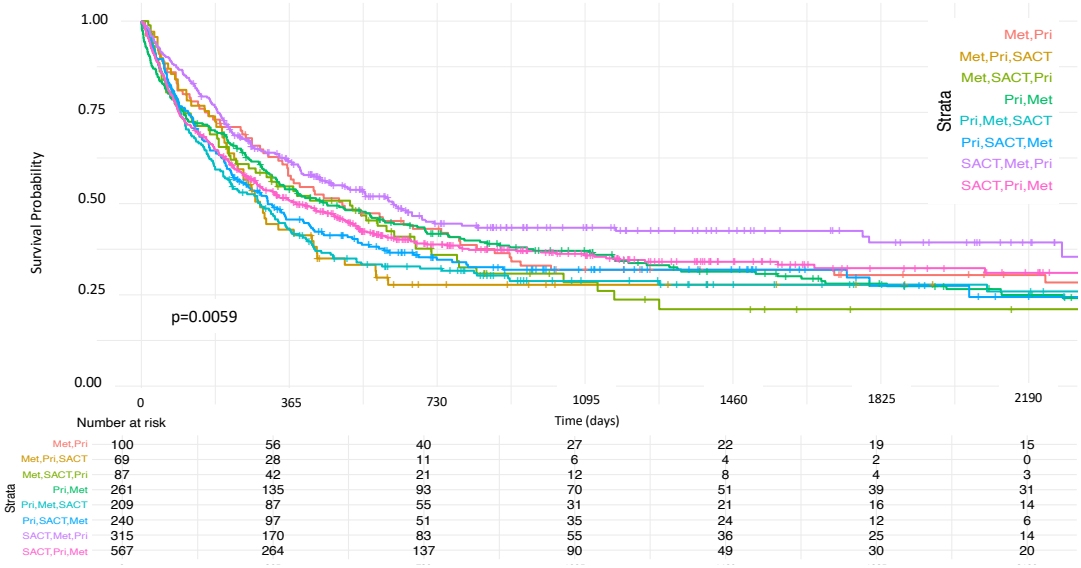


(b) Complete pathway order by MDT

Figure 5.7: Proportion of patients in each complete pathway group by (a) Cancer Alliance and (b) MDT. The y-axis represents the treating resource and the x-axis represents the complete treatment pathway group. The proportion of patients at the resource on that pathway is coloured from yellow to orange to red as the proportion increases.

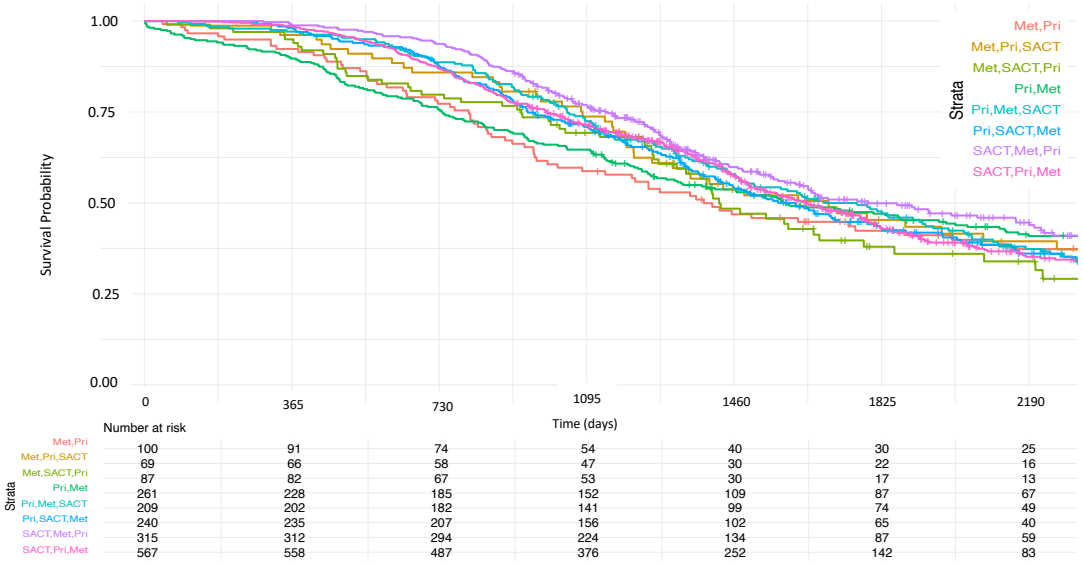


(a) Unadjusted survival from diagnosis

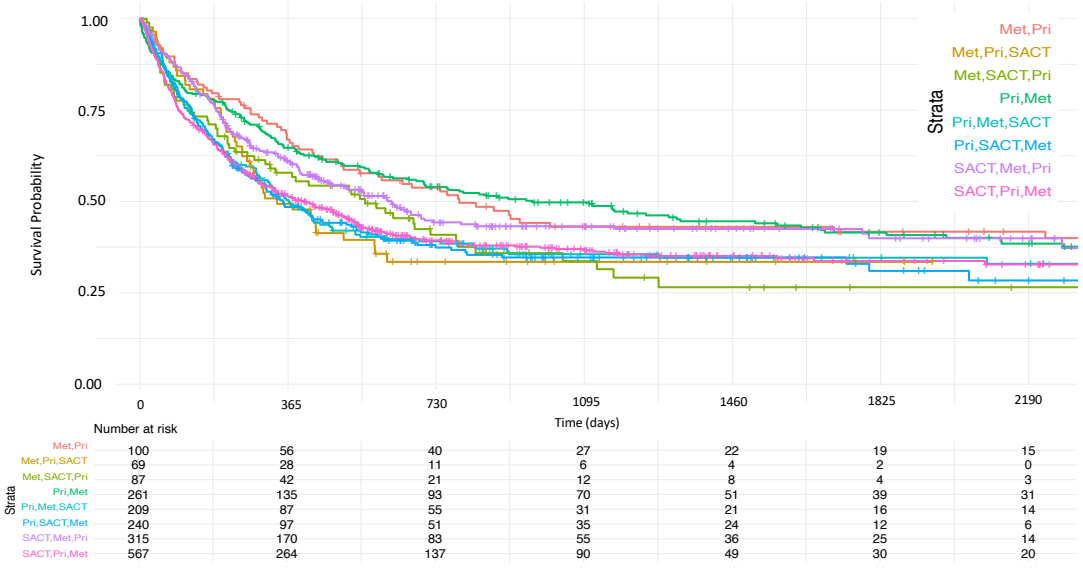


(b) Unadjusted survival from last date of treatment

Figure 5.8: Unadjusted KM curves for each of the 8 distinct, ordered complete treatment pathways with survival calculated from (a) diagnosis and (b) last date of treatment.

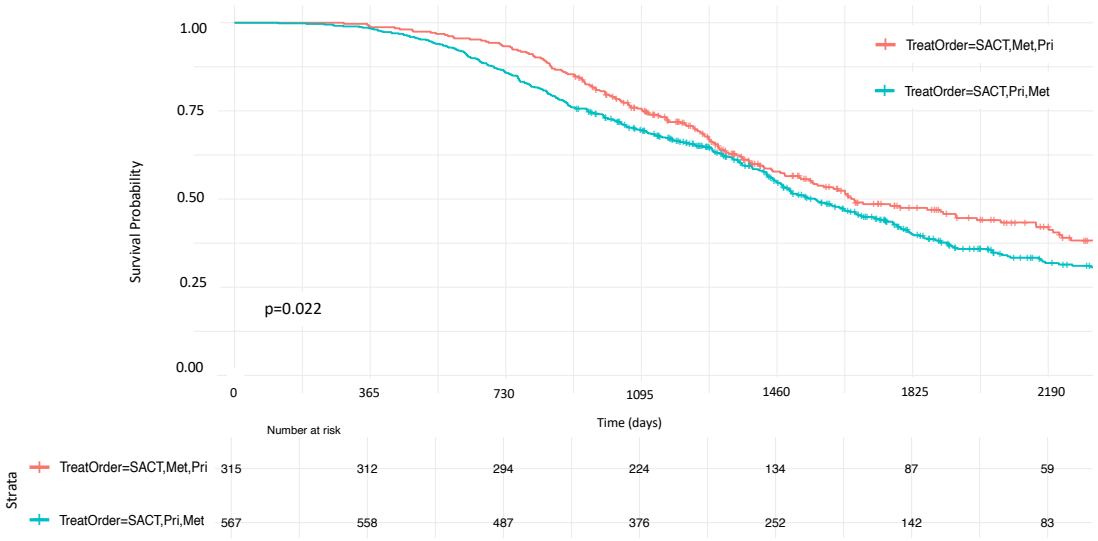


(a) Adjusted survival from diagnosis

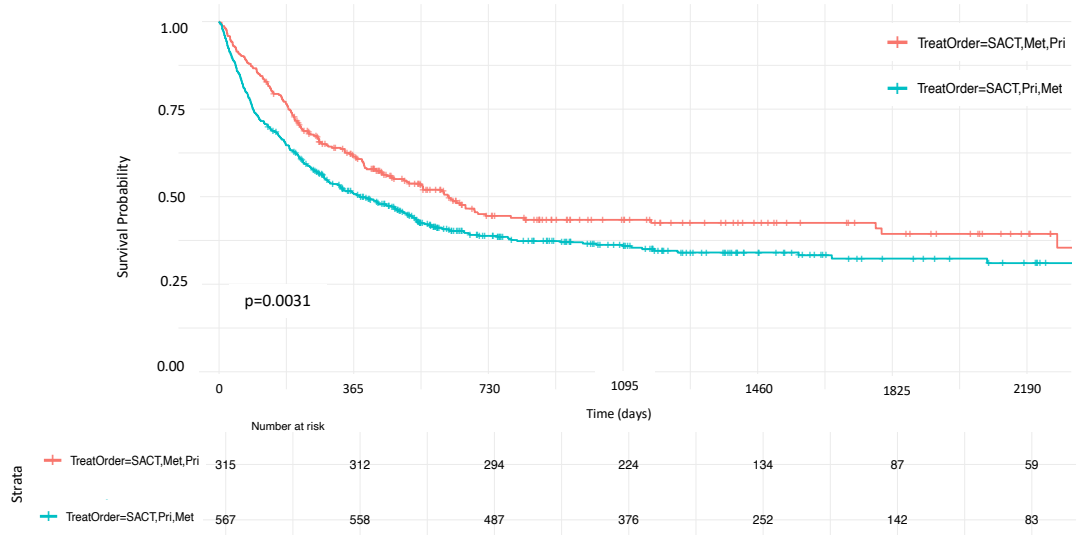


(b) Adjusted survival from last date of treatment

Figure 5.9: Adjusted KM curves for each of the 8 distinct, ordered complete treatment pathways with survival calculated from (a) diagnosis and (b) last date of treatment.

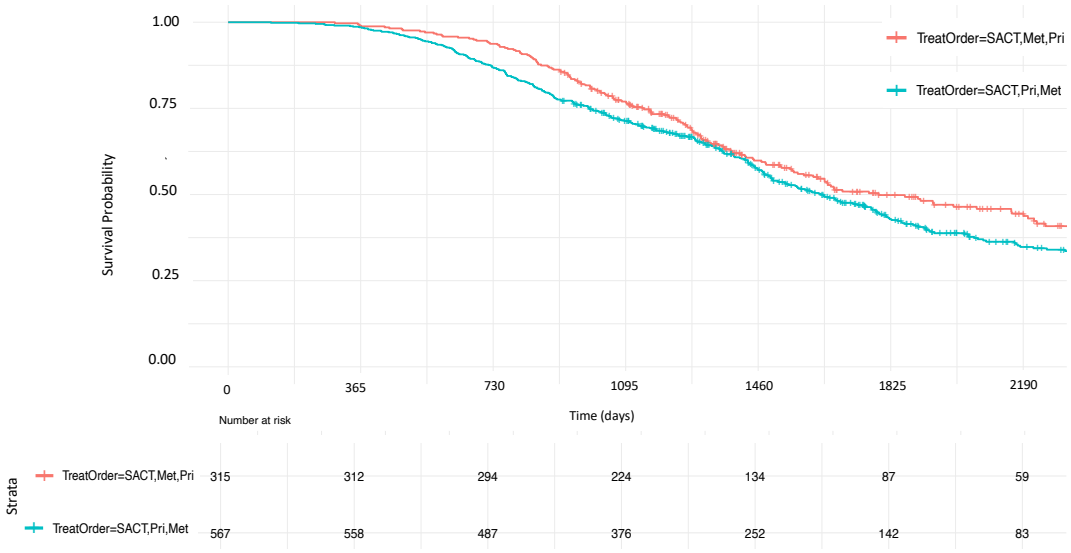


(a) Unadjusted survival from diagnosis

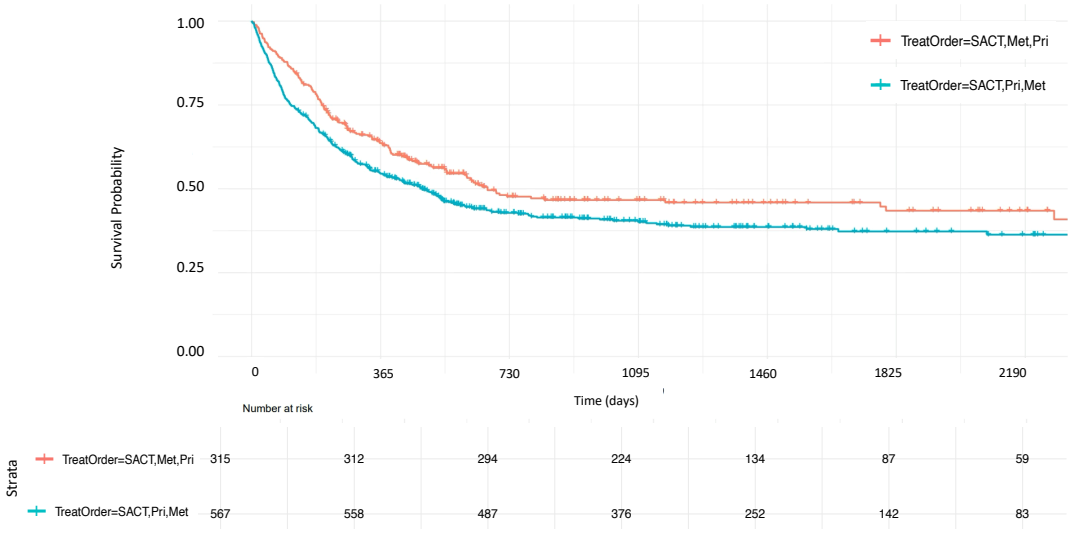


(b) Unadjusted survival from last date of treatment

Figure 5.10: Unadjusted KM curves for complete pathway groups starting with SACT with survival calculated from (a) diagnosis and (b) last date of treatment.



(a) Adjusted survival from diagnosis



(b) Adjusted survival from last date of treatment

Figure 5.11: Adjusted KM curves for complete pathway groups starting with SACT, with survival calculated from(a) diagnosis and (b) last date of treatment.

Characteristic	Both Groups (N = 882)	Group G: SACT,Met,Pri (N = 315)	Group H: SACT,Pri,Met (N = 567)
Age Group			
<= 60	430 (49%)	146 (46%)	284 (50%)
61-70	308 (35%)	121 (38%)	187 (33%)
71-80	137 (16%)	46 (15%)	91 (16%)
>80	7 (0.8%)	<5	<5
Sex			
F	259 (29%)	85 (27%)	174 (31%)
M	623 (71%)	230 (73%)	393 (69%)
Ethnicity			
White	803 (91%)	290 (92%)	513 (90%)
Other	64 (7.3%)	22 (7.0%)	42 (7.4%)
Unknown	15 (1.7%)	<5 (1.0%)	12 (2.1%)
Charlson Score			
0	801 (91%)	295 (94%)	506 (89%)
1	66 (7.5%)	18 (5.7%)	48 (8.5%)
2	10 (1.1%)	<5 (0.6%)	8 (1.4%)
3	5 (0.6%)	0 (0%)	5 (0.9%)
IMD (2015)			
1 - least deprived	206 (23%)	76 (24%)	130 (23%)
2	213 (24%)	73 (23%)	140 (25%)
3	191 (22%)	72 (23%)	119 (21%)
4	147 (17%)	53 (17%)	94 (17%)
5 - most deprived	125 (14%)	41 (13%)	84 (15%)
First SACT Regimen			
5FU/Cape +/- biological	187 (21%)	42 (13%)	145 (26%)
5FU/cape based doublet +/- biological	657 (74%)	259 (82%)	398 (70%)
Other	22 (2.5%)	7 (2.2%)	15 (2.6%)
Triplet +/- biological	16 (1.8%)	7 (2.2%)	9 (1.6%)
First Primary Treatment			
Abdominoperineal Resection	115 (14%)	47 (16%)	68 (13%)
Anterior Resection	261 (32%)	98 (33%)	163 (32%)
Exenteration	7 (0.9%)	<5 (0.3%)	6 (1.2%)
Hartmanns Resection	34 (4.2%)	19 (6.3%)	15 (3.0%)
Minor Resection	150 (19%)	58 (19%)	92 (18%)
Other Colectomy	41 (5.1%)	17 (5.7%)	24 (4.8%)
Stent	13 (1.6%)	5 (1.7%)	8 (1.6%)
Stoma	183 (23%)	55 (18%)	128 (25%)
Radiotherapy	78 (8.8%)	15 (4.8%)	63 (11%)
Number of Metastatic Sites			
1	643 (72.9%)	212 (67%)	431 (76%)
2	190 (21.5%)	78 (25%)	112 (19.7%)
3	40 (4.5%)	21 (7%)	19 (3.4%)
4	<5 (0.1%)	0 (0%)	<5 (0.2%)
Unknown	8 (0.9%)	<5 (1%)	<5 (0.7%)
Metastatic Site			
Liver	466 (53%)	197 (63%)	269 (47%)
Lung	106 (12%)	6 (2%)	100 (18%)
Oligo-Sites	231 (26%)	99 (31%)	132 (23%)
Other	71 (8.0%)	9 (3%)	62 (11%)
Unknown	8 (0.9%)	<5 (1%)	<5 (0.7%)
Stoma			
No	305 (35%)	103 (33%)	202 (36%)
Yes	577 (65%)	212 (67%)	365 (64%)

Figure 5.12: Patient characteristics for those on a complete pathway starting with a SACT regimen.

5.5 Discussion

To understand patient care faithfully researchers have to go into the details, even when it is not straightforward. The whole care pathway must be examined, as care is not received in singular instances, but as a continuous series of events. Process-informed epidemiology can accomplish this by moving analysis away from singular events and towards examining series of events.

Across England, there is variation in the proportion of stage IV rectal cancer patients completing treatment for both their primary and metastatic disease. Stage IV rectal cancer treatment journeys can be grouped into eight different treatment pathways, listed in Table 5.3. These pathways can then be grouped into three pathway types: incomplete (those who had no radical treatment), partial (those who had radical treatment to either their primary tumour or a metastatic site with or without chemotherapy), and complete (radical treatment to both the primary tumour and metastatic site with or without chemotherapy).

14% of stage IV rectal cancer patients in England diagnosed between 2012 and 2019 received a complete treatment pathway. Men, those less deprived, patients who are younger, and with less comorbidity, are more likely to have a complete pathway.

There is variation between MDTs and Cancer Alliances in the absolute proportion of patients with each pathway type and group. Risk adjusted funnel plots however indicate that when accounting for case-mix, all MDTs fall within 95% confidence bands for expected rates of patients with complete treatment. Variation in pathways is not just seen in the kinds of activities, but also in their order. There are eight different ordered pathways within the complete pathway type, with the proportion of patients on each one ranging from 4% to 31%.

There are observable and statistically significant differences in unadjusted survival between complete treatment pathways based on activity order. Significant differences were seen between pathways starting with SACT, with patients having treatment to their metastatic site before their primary site after SACT having higher survival. Patient characteristics are similar between groups, but there are differences in the sites of metastatic directed therapy.

The most significant limitation of this study is the lack of data needed to investigate further causes for differences in patient journeys. For example, having information on a patient's initial care plan will allow us to know if a difference in patient journeys occurs at the onset of care plan design, due to clinician preference at MDT or lack of access to specialist services within a Cancer Alliance, or if it is due to changes in a patient's health state as they may have side-effects or complications during progression of their treatment. Another cause could be patient preference– information that is also not routinely recorded in a EHR data–, or differences in metastasis sites– information that is not captured by the National Cancer Registry and can only be inferred by examining procedures that the patient experiences.

This lack of data also limits the possibility of investigating further if the difference in survival observed between complete treatment pathways could be due to pathway selection bias based on patient health state (such has having symptomatic primary tumour leading to treatment of the primary first), or if the pathway itself could be driving the survival differences. While a randomised control trial is needed to conclusively determine the cause, to understand the variation in more depth data on patient metastatic site(s), clinician decision making, treatment side-effects, and additional patient outcomes are required. Therefore, any conclusions drawn from the survival analysis are purely for observation and not for clinical guidance.

This study focused on pathway differences based on clinical activities. Future work will look at temporal differences between pathways to see if timings between clinical activities, or total pathway duration, are different between patient groups and if these differences have an effect on outcomes.

As far as I know this is the first study to use routine datasets to retrospectively define whole treatment pathways for stage IV rectal cancer patients. I show how to complete event-log generation from national datasets, defining treatment pathways from the event-logs, and how to assess differences between patient groups and treating hospital. Lastly, I show an initial way to link pathways to clinical outcomes– a field of research I hope continues to expand.

"Learning health systems aim to address the challenge of ensuring that the huge growth in digital health and care data can be used to improve care, to shorten the timeframe of improvement projects and ensure these are based on realworld data" –Scobie and Castle-Clarke [8]

6 Thesis Discussion

Contents

6.1 Summary of Findings 114
6.2 Limitations 116
6.3 Future Work 117

6.1 Summary of Findings

The epidemiology-informed process mining approach in Chapter 3 was able to determine variation in colon cancer treatment journeys between patient cohorts. The method is flexible and generalisable to any research question focused on understanding patient journeys using electronic health records; applying the same research question and methods to data from different EPRs showed this. The approach addresses most of the PM4H recommendations, resulting in a comprehensive method for identifying patient journeys that takes into account the specificities of healthcare and clinical processes and involves domain expertise at every stage. The findings on treatment journey variation have been presented to the Trust's equalities health board, who hope to reapply the approach to other cancer groups. The epidemiology-informed process mining approach was also used to mine the processes shown in Chapters 4 and 5.

While Chapters 3 and 5 used code-lists curated by domain experts, the activity selection and event-log generation method presented in Chapter 4 allows for the possibility of mining relevant healthcare processes, even in the absence of domain expertise, by utilising domain knowledge stored in clinical ontologies to rapidly generate and prototype code-lists. The results, however, emphasised the need to still include domain experts and end-users to ensure results are relevant and capture the breadth of local process variation, while also exploring the need to include data-driven techniques to determine events that may be exhibited in the process undetected. Chapter 4 addresses one of the recommendations for the PM4H community by explicitly defining how existing process mining concepts relate to terms and methods in the healthcare domain. The chapter shows the importance of clearly outlining which clinical activities are included in the event-log and how they were selected so that others can determine if research results are generalisable to them. This advice also applies to clinical researchers curating code-lists, as differences in code-lists will result in different conclusions being drawn about care delivery.

Chapter 5 shows that epidemiology and clinical research questions do not need to be restricted to static or singular events, but can be expanded to understand entire care processes. Understanding variation of care along an entire patient pathway revealed interesting insights into MDT-level practices and highlighted observational links between process variation and survival outcomes. Additionally, Chapter 5 highlights the strengths of the epidemiology-informed process mining approach in defining research question that leads directly to activity selection and method selection to generate a holistic process-informed epidemiology study. Framing healthcare improvement as a systems process problem represents an essential paradigm shift in how care is monitored, evaluated, and improved; once you understand how a process is implemented, you can work to change it. This includes monitoring if guidance changes issued by NICE or the NHS have been implemented, and what is their overall effect on patient outcomes and resource allocation.

6.2 Limitations

Due to technological constraints of the secure data environments, the standard process mining tools, such as ProM, were unavailable for use during the research of this thesis. This made it infeasible to use the process discovery algorithms that are often found in the process mining literature, including those developed for the healthcare context such as PALIA, as they are only implemented in ProM [119]. However, this highlighted the need for the implementation of process mining techniques to expand beyond the proprietary software to become deployable for environments such as R and STATA, which are more widely used in clinical research. The bupaR package [45] makes great strides in this way. While it does implement the heuristic miner, due to high variability of the processes mined the resulting process maps were flattened to one treatment; therefore not useful for this thesis. This limitations resulted in the usage of directly-follows graphs and dotted charts for the representation of patient journeys.

DFGs have their own limitations, however, as interpretation of pathways can be misleading as information on the graphs is conditional on one activity following another. Additionally, processes with flexible ordering and loops can be difficult to read and may require additional filtering, oversimplifying the graphs [43]. While the DFGs were useful in providing clear information for this thesis, others wishing to use these methods on different research questions and datasets may not find them applicable.

The most significant limitation of this thesis is the problem of data depth. The disparity between what is recorded in a care record, how it is transformed into secondary research datasets, and what is actually required for research into care delivery is large. For this thesis, the biggest challenge was in not having recorded metastatic sites for stage IV rectal cancer patients, even though by definition stage IV patients have metastasis sites, the National Cancer Registry Service does not systematically record them. Care plans are also not routinely recorded; if they are, they are often not easily retrievable from EPR systems.

This situation is likely to improve over time, with the increased use of care records for planning, research, and development. Techniques such as epidemiology-informed process mining can be used to help identify and prioritise data points for inclusion in future care records and reporting datasets.

6.3 Future Work

Beyond the specific future work listed in each chapter's discussion, there are two areas for future work that could drive forward epidemiology and process mining for the understanding of patient journeys: the formation of process-informed EPR systems and specialised conformance checking methods.

Process-informed EPR systems can address the limitation of data depth needed to analyse care delivery faithfully. The required redesign of existing clinical systems and datasets can occur by focusing on an element of the virtuous cycle of a learning health system: turning practice into data. This includes the standardised documentation of patient care plans, clinical guidelines, core care plans, clinician decision making, the direct actions on a patient's journey, and the results of those actions; the metadata (who recorded the data point, when, and where) is also required for effective secondary data usage. EPR systems should allow for iterative design as new research questions will require additional information to be systematically capture. openEHRs work to create standards for clinical models and software to ensure interoperability between solutions is a good starting point for standardising data collection formats [120].

Once this data is captured, effective methods must be available to compare intended plans and realised journeys. Current conformance checking methods are not suited to healthcare due to the warranted, and even encouraged, variability of patient journeys due to personalisation of care. Having hard "rules" for identifying deviations from a care plan could result in the over-penalisation of clinicians and hospitals. This is already an issue as the NHS has many targets, which if missed can result in financial penalties for the hospital. Future conformance checking methods should seek to imbed contextual data which can indicate if deviations

are planned and clinically valid and therefore should not be penalised, as shown by the manual conformance checking analysis in Section 3.3.4.5. Methods could also include using confidence intervals for determining unwanted process variability, where only hospitals who exceed a threshold for variation are flagged.

Concurrent to the development of healthcare specific conformance checking techniques, as well as process-informed EPRs, is the need for computable and machine readable guidelines. This work is currently underway with initiatives such as the World Health Organisation's digital adaptation kits [55], but should remain closely linked to the development of conformance checking methods and EPR system design to ensure compatibility.

It is my hope that with the development of EPR systems and datasets that capture patient journeys more faithfully, and the closer integration of epidemiology and process mining techniques, that a learning health system can be realised in the NHS to allow for real-time monitoring of care delivery to ensure standardised, high quality care for all.

Appendices

A
Additional DFGs

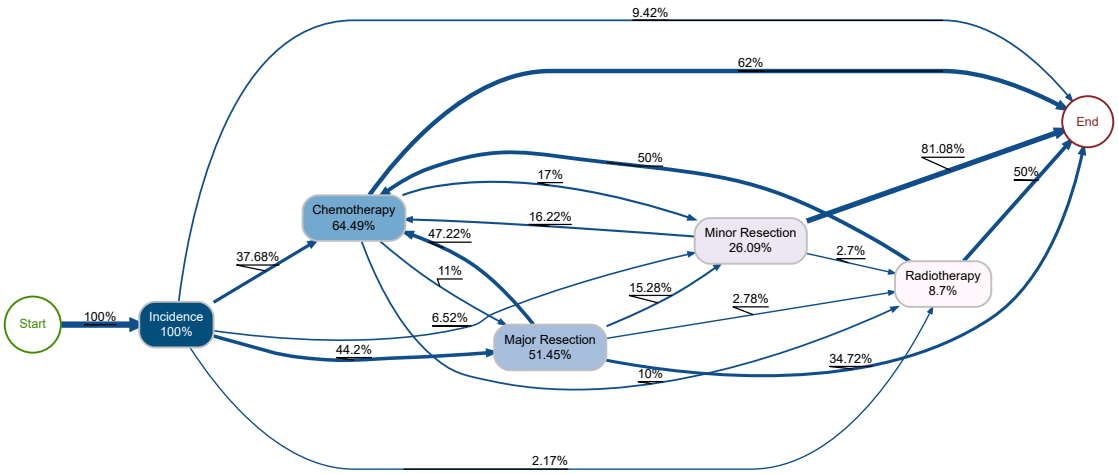


Figure A.1: DFG showing treatment journeys for colon cancer patients aged under 49 at incidence.

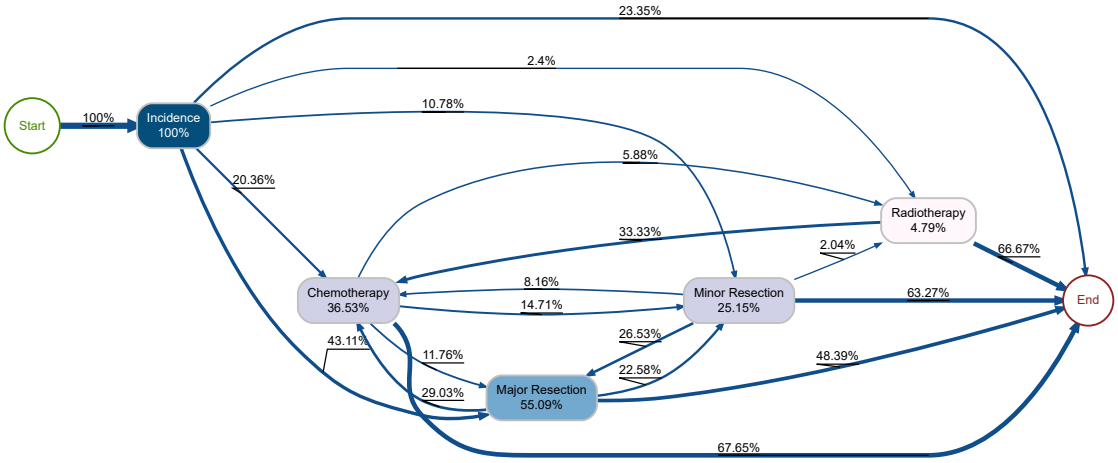


Figure A.2: DFG showing treatment journeys for colon cancer patients living in a post-code with an IMD score of 3 at incidence.

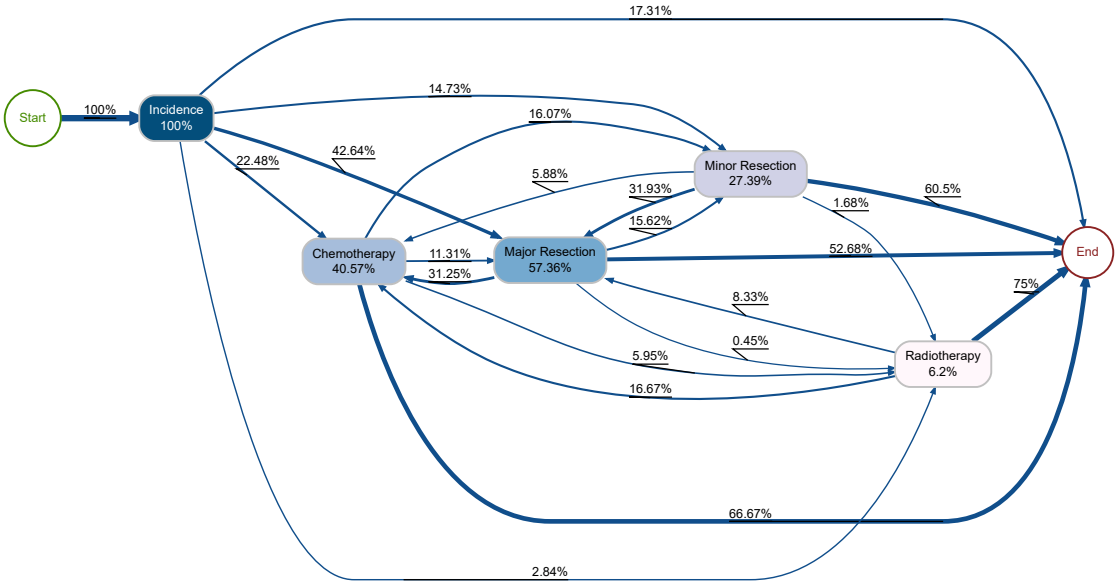


Figure A.3: DFG showing treatment journeys for colon cancer patients living in a post-code with an IMD score of 4 at incidence.

B

Patient Demographics by Cancer Alliance

Characteristic	Overall N = 12,826	Cheshire and Merseyside Cancer Alliance N = 682	East Midlands Cancer Alliance N = 937	East of England - North Cancer Alliance N = 844	East of England - South Cancer Alliance N = 707	Greater Manchester Cancer Alliance N = 688	Humber, Coast and Vale Cancer Alliance N = 357	Kent and Medway Cancer Alliance N = 426	Lancashire & South Cumbria Cancer Alliance N = 408	North Central London Cancer Alliance N = 263	North East London Cancer Alliance N = 271	Northern Cancer Alliance N = 880	Peninsula Cancer Alliance N = 519	Somerset, Wiltshire, Avon and Gloucestershire Cancer Alliance N = 707	South East London Cancer Alliance N = 245	South Yorkshire & Bassetlaw Cancer Alliance N = 438	Surrey and Sussex Cancer Alliance N = 670	Thames Valley Cancer Alliance N = 602	Wessex Cancer Alliance N = 482	West London Cancer Alliance N = 575	West Midlands Cancer Alliance N = 1,521	West Yorkshire and Harrogate Cancer Alliance N = 604
age_group																						
<=60	3,427 (27%)	188 (28%)	257 (27%)	191 (23%)	190 (27%)	201 (29%)	78 (22%)	107 (25%)	105 (26%)	90 (34%)	84 (31%)	208 (24%)	123 (24%)	197 (28%)	84 (34%)	117 (27%)	175 (26%)	175 (29%)	122 (25%)	180 (31%)	402 (26%)	153 (25%)
61-70	3,344 (26%)	177 (26%)	246 (26%)	249 (30%)	184 (26%)	193 (28%)	99 (28%)	105 (25%)	109 (27%)	70 (27%)	69 (25%)	220 (25%)	113 (22%)	180 (25%)	54 (22%)	111 (25%)	164 (24%)	160 (27%)	126 (26%)	155 (27%)	397 (26%)	163 (27%)
71-80	3,554 (28%)	207 (30%)	252 (27%)	213 (25%)	185 (26%)	202 (29%)	128 (36%)	120 (28%)	107 (26%)	56 (21%)	68 (25%)	260 (30%)	168 (32%)	188 (27%)	57 (23%)	118 (27%)	195 (29%)	146 (24%)	132 (27%)	124 (22%)	442 (29%)	186 (31%)
>80	2,501 (19%)	110 (16%)	182 (19%)	191 (23%)	148 (21%)	92 (13%)	52 (15%)	94 (22%)	87 (21%)	47 (18%)	50 (18%)	192 (22%)	115 (22%)	142 (20%)	50 (20%)	92 (21%)	136 (20%)	121 (20%)	102 (21%)	116 (20%)	280 (18%)	102 (17%)
Sex																						
F	4,440 (35%)	230 (34%)	318 (34%)	289 (34%)	243 (34%)	226 (33%)	123 (34%)	173 (41%)	149 (37%)	116 (44%)	87 (32%)	285 (32%)	192 (37%)	248 (35%)	86 (35%)	149 (34%)	241 (36%)	214 (36%)	170 (35%)	213 (37%)	483 (32%)	205 (34%)
M	8,386 (65%)	452 (66%)	619 (66%)	555 (66%)	464 (66%)	462 (67%)	234 (66%)	253 (59%)	259 (63%)	147 (56%)	184 (68%)	595 (68%)	327 (63%)	459 (65%)	159 (65%)	289 (66%)	429 (64%)	388 (64%)	312 (65%)	362 (63%)	1,038 (68%)	399 (66%)
Met Site																						
Liver	1,439 (11%)	88 (13%)	96 (10%)	103 (12%)	59 (8.3%)	72 (10%)	38 (11%)	53 (12%)	31 (7.6%)	44 (17%)	26 (9.6%)	81 (9.2%)	43 (8.3%)	79 (11%)	29 (12%)	36 (8.2%)	85 (13%)	76 (13%)	69 (14%)	76 (13%)	193 (13%)	62 (10%)
Lung	486 (3.8%)	23 (3.4%)	29 (3.1%)	37 (4.4%)	24 (3.4%)	23 (3.3%)	8 (2.2%)	18 (4.2%)	13 (3.2%)	12 (4.6%)	8 (3.0%)	36 (4.1%)	15 (2.9%)	29 (4.1%)	18 (7.3%)	11 (2.5%)	33 (4.9%)	25 (4.2%)	17 (3.5%)	26 (4.5%)	59 (3.9%)	22 (3.6%)
OligoSites	567 (4.4%)	36 (5.3%)	45 (4.8%)	40 (4.7%)	34 (4.8%)	37 (5.4%)	9 (2.5%)	15 (3.5%)	16 (3.9%)	11 (4.2%)	9 (3.3%)	24 (2.7%)	30 (5.8%)	29 (4.1%)	14 (5.7%)	11 (2.5%)	27 (4.0%)	32 (5.3%)	21 (4.4%)	35 (6.1%)	74 (4.9%)	18 (3.0%)
Other	374 (2.9%)	13 (1.9%)	25 (2.7%)	20 (2.4%)	20 (2.8%)	22 (3.2%)	7 (2.0%)	11 (2.6%)	13 (3.2%)	5 (1.9%)	9 (3.3%)	26 (3.0%)	17 (3.3%)	23 (3.3%)	9 (3.7%)	16 (3.7%)	18 (2.7%)	16 (2.7%)	15 (3.1%)	25 (4.3%)	52 (3.4%)	12 (2.0%)
Unknown	9,960 (78%)	522 (77%)	742 (79%)	644 (76%)	570 (81%)	534 (78%)	295 (83%)	329 (77%)	335 (82%)	191 (73%)	219 (81%)	713 (81%)	414 (80%)	547 (77%)	175 (71%)	364 (83%)	507 (76%)	453 (75%)	360 (75%)	413 (72%)	1,143 (75%)	490 (81%)
Ethnicity																						
Other	665 (5.2%)	13 (1.9%)	50 (5.3%)	23 (2.7%)	31 (4.4%)	40 (5.8%)	<5 (0.6%)	12 (2.8%)	12 (2.9%)	59 (22%)	68 (25%)	<5 (0.3%)	<5 (0.4%)	8 (1.1%)	39 (16%)	9 (2.1%)	18 (2.7%)	25 (4.2%)	14 (2.9%)	125 (22%)	87 (5.7%)	25 (4.1%)
Unknown	591 (4.6%)	5 (0.7%)	52 (5.5%)	48 (5.7%)	41 (5.8%)	8 (1.2%)	20 (5.6%)	8 (1.9%)	18 (4.4%)	7 (2.7%)	6 (2.2%)	37 (4.2%)	13 (2.5%)	58 (8.2%)	20 (8.2%)	5 (1.1%)	57 (8.5%)	76 (13%)	20 (4.1%)	18 (3.1%)	63 (4.1%)	11 (1.8%)
White	11,570 (90%)	664 (97%)	835 (89%)	773 (92%)	635 (90%)	640 (93%)	335 (94%)	406 (95%)	378 (93%)	197 (75%)	197 (73%)	840 (95%)	504 (97%)	641 (91%)	186 (76%)	424 (97%)	595 (89%)	501 (83%)	448 (93%)	432 (75%)	1,371 (90%)	568 (94%)
charlson1yrcat																						
0	10,084 (79%)	543 (80%)	731 (78%)	687 (81%)	563 (80%)	508 (74%)	273 (76%)	332 (78%)	309 (76%)	188 (71%)	192 (71%)	658 (75%)	420 (81%)	576 (81%)	193 (79%)	351 (80%)	562 (84%)	490 (81%)	389 (81%)	449 (78%)	1,196 (79%)	474 (78%)
1	1,627 (13%)	83 (12%)	124 (13%)	85 (10%)	89 (13%)	110 (16%)	54 (15%)	61 (14%)	60 (15%)	46 (17%)	47 (17%)	131 (15%)	51 (9.8%)	69 (9.8%)	28 (11%)	56 (13%)	61 (9.1%)	61 (10%)	56 (12%)	80 (14%)	192 (13%)	83 (14%)
2	609 (4.7%)	33 (4.8%)	49 (5.2%)	44 (5.2%)	26 (3.7%)	39 (5.7%)	15 (4.2%)	21 (4.9%)	22 (5.4%)	19 (7.2%)	22 (8.1%)	47 (5.3%)	19 (3.7%)	39 (5.5%)	14 (5.7%)	16 (3.7%)	17 (2.5%)	24 (4.0%)	19 (3.9%)	25 (4.3%)	71 (4.7%)	28 (4.6%)
3	506 (3.9%)	23 (3.4%)	33 (3.5%)	28 (3.3%)	29 (4.1%)	31 (4.5%)	15 (4.2%)	12 (2.8%)	17 (4.2%)	10 (3.8%)	10 (3.7%)	44 (5.0%)	29 (5.6%)	23 (3.3%)	10 (4.1%)	15 (3.4%)	30 (4.5%)	27 (4.5%)	18 (3.7%)	21 (3.7%)	62 (4.1%)	19 (3.1%)
imd2015																						
1 - least deprived	2,488 (19%)	134 (20%)	167 (18%)	161 (19%)	153 (22%)	101 (15%)	59 (17%)	89 (21%)	71 (17%)	25 (9.5%)	13 (4.8%)	140 (16%)	44 (8.5%)	190 (27%)	39 (16%)	62 (14%)	201 (30%)	244 (41%)	120 (25%)	111 (19%)	230 (15%)	134 (22%)
2	2,736 (21%)	131 (19%)	231 (25%)	227 (27%)	175 (25%)	115 (17%)	85 (24%)	89 (21%)	93 (23%)	26 (9.9%)	21 (7.7%)	140 (16%)	128 (25%)	186 (26%)	41 (17%)	76 (17%)	170 (25%)	140 (23%)	124 (26%)	106 (18%)	310 (20%)	122 (20%)
3	2,678 (21%)	109 (16%)	178 (19%)	229 (27%)	178 (25%)	123 (18%)	84 (24%)	104 (24%)	80 (20%)	59 (22%)	34 (13%)	140 (16%)	161 (31%)	149 (21%)	48 (20%)	72 (16%)	167 (25%)	108 (18%)	109 (23%)	148 (26%)	301 (20%)	97 (16%)
4	2,545 (20%)	99 (15%)	192 (20%)	145 (17%)	133 (19%)	124 (18%)	60 (17%)	90 (21%)	75 (18%)	79 (30%)	107 (39%)	208 (24%)	113 (22%)	124 (18%)	58 (24%)	106 (24%)	98 (15%)	81 (13%)	96 (20%)	136 (24%)	311 (20%)	110 (18%)
5 - most deprived	2,379 (19%)	209 (31%)	169 (18%)	82 (9.7%)	68 (9.6%)	225 (33%)	69 (19%)	54 (13%)	89 (22%)	74 (28%)	96 (35%)	252 (29%)	73 (14%)	58 (8.2%)	59 (24%)	122 (28%)	34 (5.1%)	29 (4.8%)	33 (6.8%)	74 (13%)	369 (24%)	141 (23%)
Outcome																						
0	10,979 (86%)	567 (83%)	804 (86%)	717 (85%)	629 (89%)	581 (84%)	311 (87%)	380 (89%)	366 (90%)	219 (83%)	237 (87%)	778 (88%)	447 (86%)	597 (84%)	201 (82%)	393 (90%)	552 (82%)	519 (86%)	388 (80%)	474 (82%)	1,294 (85%)	525 (87%)
1	1,848 (14%)	115 (17%)	133 (14%)	127 (15%)	78 (11%)	107 (16%)	46 (13%)	46 (11%)	42 (10%)	44 (17%)	34 (13%)	102 (12%)	72 (14%)	110 (16%)	44 (18%)	45 (10%)	119 (18%)	83 (14%)	94 (20%)	101 (18%)	227 (15%)	79 (13%)
Treatment Type																						
NoTreatment	5,052 (39%)	262 (38%)	354 (38%)	341 (40%)	307 (43%)	246 (36%)	142 (40%)	216 (51%)	141 (35%)	88 (33%)	104 (38%)	398 (45%)	191 (37%)	286 (40%)	72 (29%)	207 (47%)	257 (38%)	202 (34%)	178 (37%)	222 (39%)	625 (41%)	213 (35%)
PartialTreatment	5,927 (46%)	305 (45%)	450 (48%)	376 (45%)	322 (46%)	335 (49%)	169 (47%)	164 (38%)	225 (55%)	131 (50%)	133 (49%)	380 (43%)	256 (49%)	311 (44%)	129 (53%)	186 (42%)	295 (44%)	317 (53%)	210 (44%)	252 (44%)	669 (44%)	312 (52%)
CompleteTreatment	1,848 (14%)	115 (17%)	133 (14%)	127 (15%)	78 (11%)	107 (16%)	46 (13%)	46 (11%)	42 (10%)	44 (17%)	34 (13%)	102 (12%)	72 (14%)	110 (16%)	44 (18%)	45 (10%)	119 (18%)	83 (14%)	94 (20%)	101 (18%)	227 (15%)	79 (13%)
Survival (days)	396 (122, 909)	390 (124, 870)	367 (121, 907)	353 (108, 854)	412 (139, 828)	392 (134, 955)	407 (135, 869)	384 (101, 803)	360 (108, 850)	459 (129, 1,010)	411 (117, 1,002)	340 (112, 780)	388 (118, 847)	387 (116, 908)	551 (146, 1,120)	386 (114, 810)	494 (132, 1,007)	397 (124, 896)	427 (142, 976)	479 (135, 1,075)	370 (107, 908)	434 (153, 956)

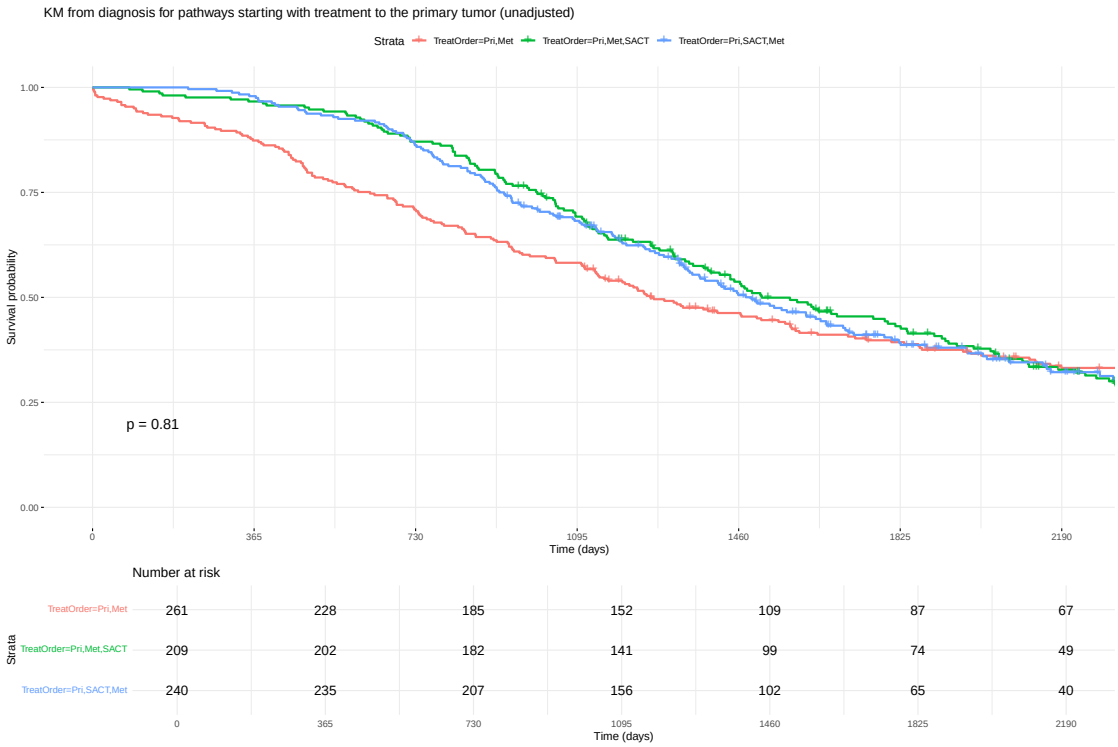
C

Patient Demographics by first SACT regimen

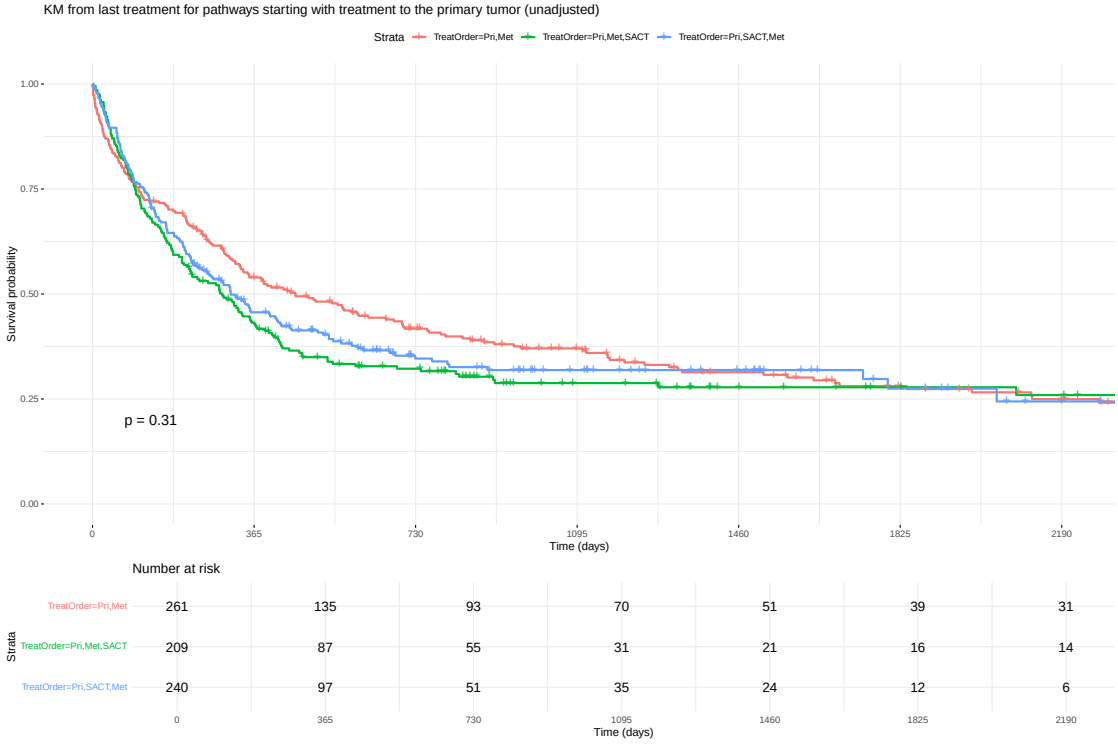
Characteristic	Overall N = 6,144	5FU/Cape +/- biological N = 1,167	5FU/cape based doublet +/- biological N = 4,541	Other N=3,38	Triplet +/- biological N = 98
Age	65 (56, 72)	70 (60, 77)	64 (55, 71)	63 (53, 72)	55 (47, 62)
Sex					
F	1,949 (32%)	388 (33%)	1,405 (31%)	122 (36%)	34 (35%)
M	4,195 (68%)	779 (67%)	3,136 (69%)	216 (64%)	64 (65%)
Ethnicity					
White	5,545 (90%)	1,076 (92%)	4,071 (90%)	316 (93%)	82 (84%)
Other	364 (5.9%)	54 (4.6%)	285 (6.3%)	15 (4.4%)	10 (10%)
Unknown	235 (3.8%)	37 (3.2%)	185 (4.1%)	7 (2.1%)	6 (6.1%)
Charlson Score					
0	5,354 (87%)	1,008 (86%)	3,978 (88%)	277 (82%)	91 (93%)
1	584 (9.5%)	114 (9.8%)	428 (9.4%)	36 (11%)	6 (6.1%)
2	147 (2.4%)	29 (2.5%)	103 (2.3%)	14 (4.1%)	<5 (1.0%)
3	59 (1.0%)	16 (1.4%)	32 (0.7%)	11 (3.3%)	0 (0%)
IMD (2015)					
1 - least deprived	1,289 (21%)	225 (19%)	967 (21%)	81 (24%)	16 (16%)
2	1,376 (22%)	262 (22%)	1,020 (22%)	71 (21%)	23 (23%)
3	1,323 (22%)	239 (20%)	995 (22%)	64 (19%)	25 (26%)
4	1,127 (18%)	219 (19%)	821 (18%)	65 (19%)	22 (22%)
5 - most deprived	1,029 (17%)	222 (19%)	738 (16%)	57 (17%)	12 (12%)

D

KM Curves for complete pathways starting with primary site directed therapy

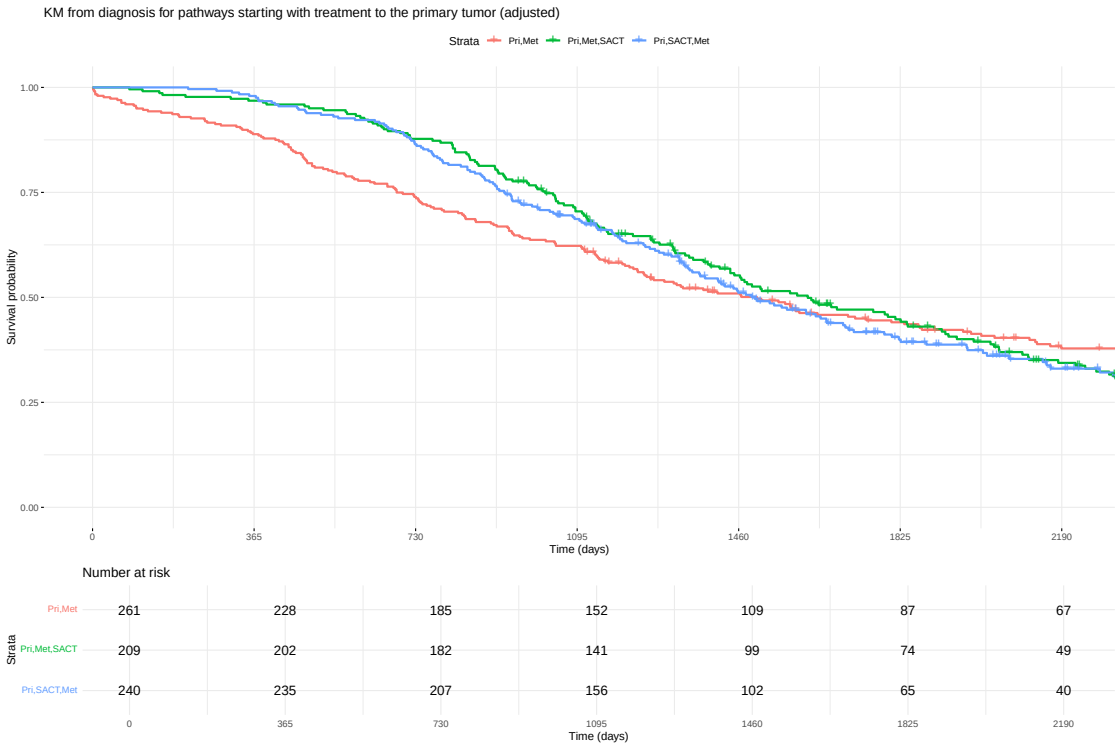


(a) Unadjusted survival from diagnosis

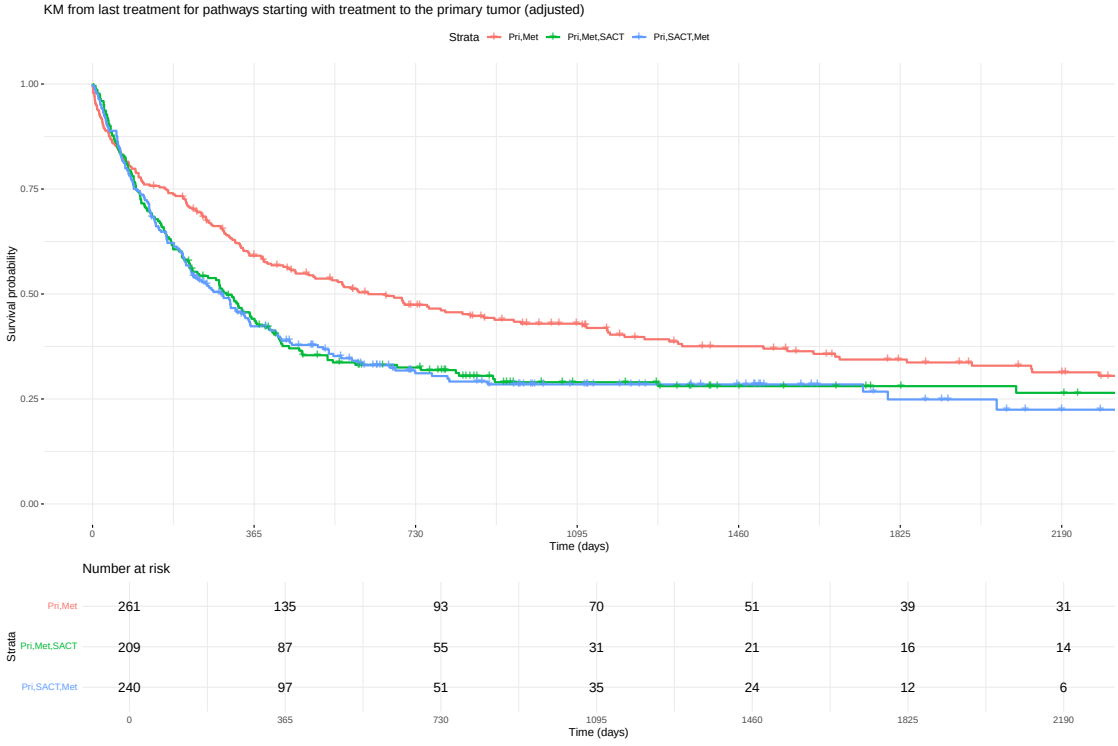


(b) Unadjusted survival from last date of treatment

Figure D.1: Unadjusted KM curves for complete pathway Groups A,B,C starting with primary site directed therapy, with survival calculated from (a) diagnosis and (b) last date of treatment.



(a) Adjusted survival from diagnosis

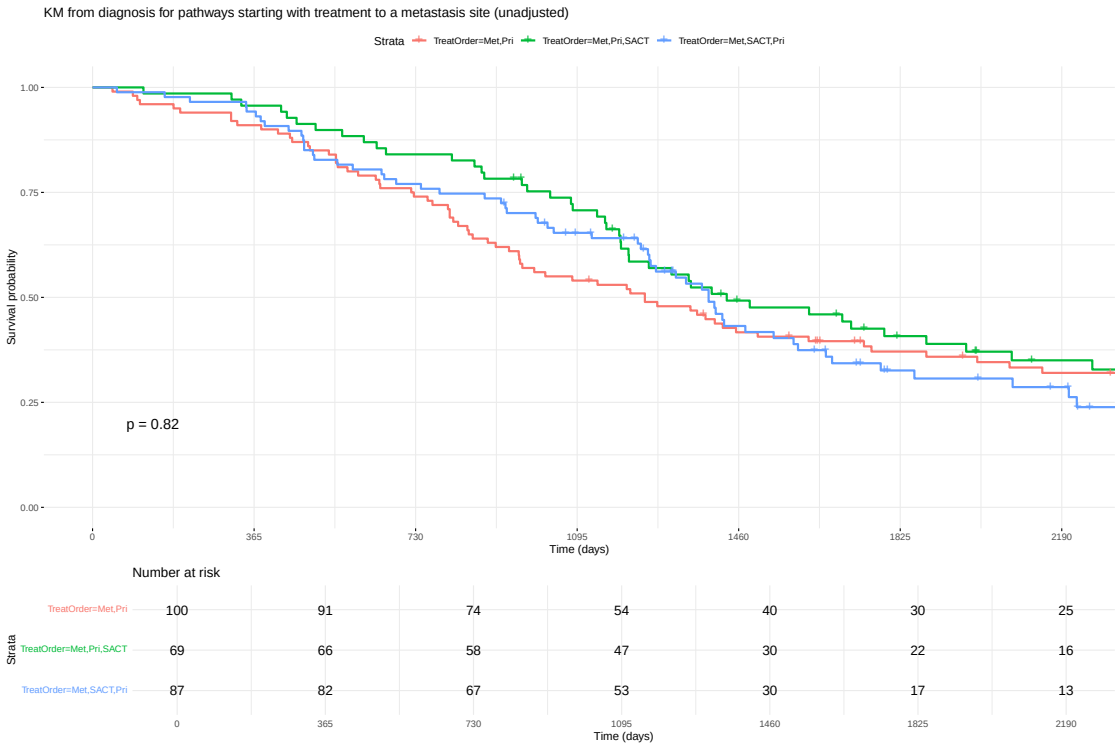


(b) Adjusted survival from last date of treatment

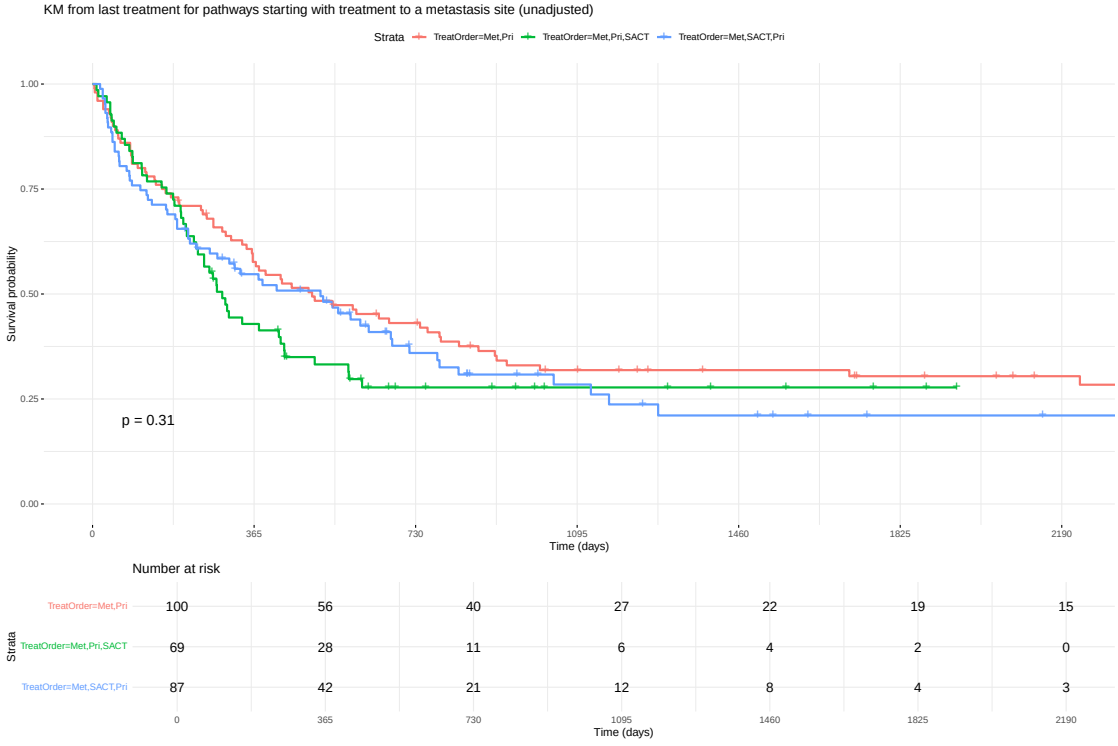
Figure D.2: Adjusted KM curves for complete pathway Groups A,B,C starting with primary site directed therapy, with survival calculated from (a) diagnosis and (b) last date of treatment.

E

KM Curves for complete pathways starting with metastasis directed therapy

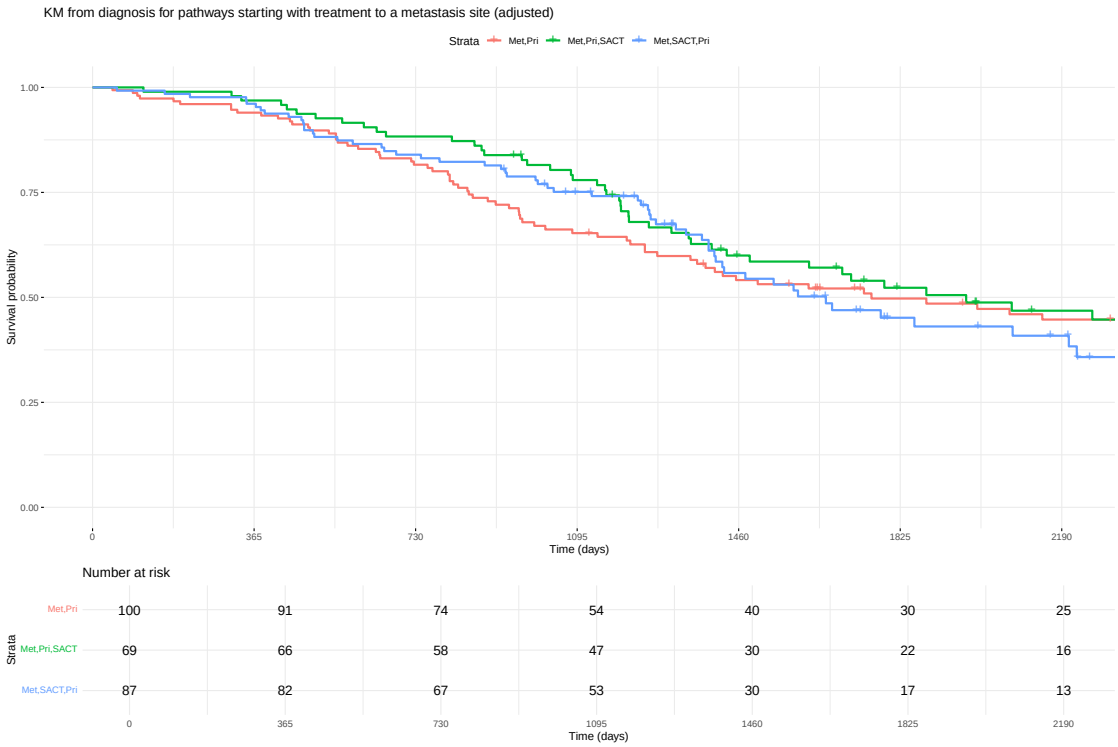


(a) Unadjusted survival from diagnosis

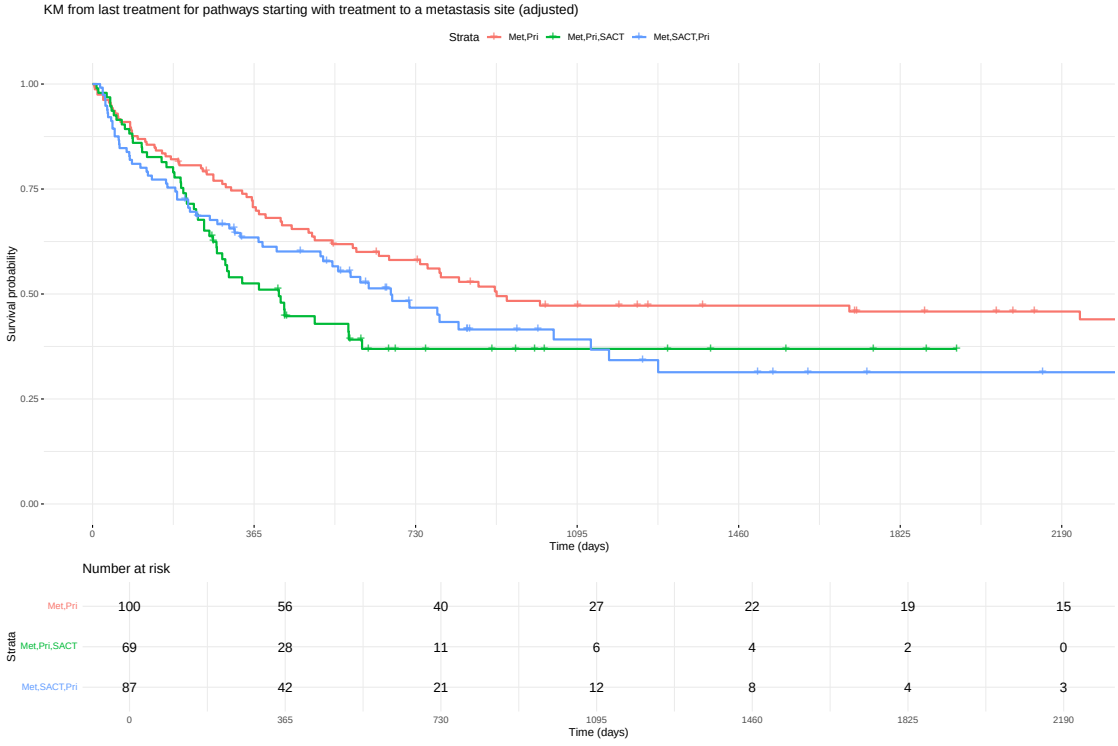


(b) Unadjusted survival from last date of treatment

Figure E.1: Unadjusted KM curves for complete pathway Groups A,B,C starting with metastasis site directed therapy, with survival calculated from (a) diagnosis and (b) last date of treatment.



(a) Adjusted survival from diagnosis



(b) Adjusted survival from last date of treatment

Figure E.2: Adjusted KM curves for complete pathway Groups A,B,C starting with metastasis site directed therapy, with survival calculated from (a) diagnosis and (b) last date of treatment.

References

[1] S. Benitez Majano, C. Di Girolamo, B. Rachet, *et al.*, "Surgical treatment and survival from colorectal cancer in Denmark, England, Norway, and Sweden: A population-based study," *The Lancet Oncology*, vol. 20, no. 1, pp. 74–87, Jan. 2019. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S1470204518306466 (visited on 12/15/2024).

[2] S. Ling, M.-A. Luque Fernandez, M. Quaresma, A. Belot, and B. Rachet, "Inequalities in treatment among patients with colon and rectal cancer: A multistate survival model using data from England national cancer registry 2012–2016," *British Journal of Cancer*, vol. 130, no. 1, pp. 88–98, Jan. 2024. [Online]. Available: https://www.nature.com/articles/s41416-023-02440-6 (visited on 12/14/2024).

[3] E. J. A. Morris, R. Goldacre, E. Spata, *et al.*, "Impact of the COVID-19 pandemic on the detection and management of colorectal cancer in England: A population-based study," *The Lancet Gastroenterology & Hepatology*, vol. 6, no. 3, pp. 199–208, Mar. 2021. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S2468125321000054 (visited on 06/10/2022).

[4] K. Spencer, C. M. Jones, R. Girdler, *et al.*, "The impact of the COVID-19 pandemic on radiotherapy services in England, UK: A population-based study," *The Lancet Oncology*, vol. 22, no. 3, pp. 309–320, Mar. 2021. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S1470204520307439 (visited on 06/10/2022).

[5] T. Foley and L. Vale, "A framework for understanding, designing, developing and evaluating learning health systems," *Learning Health Systems*, vol. n/a, no. n/a, e10315, May 1, 2022. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1002/lrh2.10315 (visited on 06/10/2022).

[6] World Health Organization and Alliance for Health Policy and Systems Research, "Learning health systems: Pathways to progress: Flagship report of the Alliance for Health Policy and Systems Research," World Health Organization, Geneva, 2021. [Online]. Available: https://apps.who.int/iris/handle/10665/344891 (visited on 06/29/2022).

[7] Dr Tom Foley and Dr Fergus Fairmichael, "The Potential of Learning Health Systems," The Learning Healthcare Project, Nov. 2015.

[8] D. S. Scobie and S. Castle-Clarke, "What can the NHS learn from learning health systems?" Nuffield Trust, May 2019.

[9] W. van der Aalst, *Process Mining: Data Science in Action*. Springer, 2016, vol. 2.

[10] 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," in *Process Support and Knowledge Representation in Health Care*, R. Lenz, S. Miksch, M. Peleg, M. Reichert, D. Riaño, and A. ten Teije, Eds., red. by D. Hutchison, T. Kanade, J. Kittler, *et al.*, vol. 7738, Berlin, Heidelberg: Springer Berlin Heidelberg, 2013, pp. 140–153. [Online]. Available: http://link.springer.com/10.1007/978-3-642-36438-9_10 (visited on 06/02/2022).

[11] W. van der Aalst, A. Adriansyah, A. K. A. de Medeiros, *et al.*, "Process Mining Manifesto," in *Business Process Management Workshops*, F. Daniel, K. Barkaoui, and S. Dustdar, Eds., vol. 99, Berlin, Heidelberg: Springer Berlin Heidelberg, 2012, pp. 169–194. [Online]. Available: http://link.springer.com/10.1007/978-3-642-28108-2_19 (visited on 05/28/2022).

[12] E. Rojas, J. Munoz-Gama, M. Sepúlveda, and D. Capurro, "Process mining in healthcare: A literature review," *Journal of Biomedical Informatics*, vol. 61, pp. 224–236, Jun. 2016. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S1532046416300296 (visited on 02/21/2022).

[13] E. De Roock and N. Martin, "Process mining in healthcare – An updated perspective on the state of the art," *Journal of Biomedical Informatics*, vol. 127, p. 103 995, Mar. 2022. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S1532046422000119 (visited on 05/31/2022).

[14] A. P. Kurniati, O. Johnson, D. Hogg, and G. Hall, "Process mining in oncology: A literature review," in *2016 6th International Conference on Information Communication and Management (ICICM)*, Oct. 2016, pp. 291–297.

[15] J. Munoz-Gama, N. Martin, C. Fernandez-Llatas, *et al.*, "Process mining for healthcare: Characteristics and challenges," *Journal of Biomedical Informatics*, vol. 127, p. 103 994, Mar. 2022. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S1532046422000107 (visited on 05/28/2022).

[16] E. J. A. Morris, D. Forman, J. D. Thomas, *et al.*, "Surgical management and outcomes of colorectal cancer liver metastases," *British Journal of Surgery*, vol. 97, no. 7, pp. 1110–1118, Jul. 1, 2010. [Online]. Available: https://doi.org/10.1002/bjs.7032 (visited on 03/15/2023).

[17] E. Morris, P. Quirke, J. D. Thomas, L. Fairley, B. Cottier, and D. Forman, "Unacceptable variation in abdominoperineal excision rates for rectal cancer: Time to intervene?" *Gut*, vol. 57, no. 12, pp. 1690–1697, Jul. 30, 2008. [Online]. Available: https://gut.bmj.com/lookup/doi/10.1136/gut.2007.137877 (visited on 03/13/2023).

[18] Olivia Cavian, Penny Dash, Jean Drouin, Tim Fountaine, and Farhad Riahi, "Using care pathways to improve health systems," Health International, McKinsey Co., 11, 2011.

[19] E. Aspland, D. Gartner, and P. Harper, "Clinical pathway modelling: A literature review," *Health Systems*, vol. 10, no. 1, pp. 1–23, Jan. 2, 2021. [Online]. Available: https://doi.org/10.1080/20476965.2019.1652547 (visited on 08/02/2021).

[20] A. K. Lawal, T. Rotter, L. Kinsman, *et al.*, "What is a clinical pathway? Refinement of an operational definition to identify clinical pathway studies for a Cochrane systematic review," *BMC Medicine*, vol. 14, no. 1, p. 35, Feb. 23, 2016. [Online]. Available: https://doi.org/10.1186/s12916-016-0580-z (visited on 05/31/2022).

[21] M. Cho, K. Kim, J. Lim, *et al.*, "Developing data-driven clinical pathways using electronic health records: The cases of total laparoscopic hysterectomy and rotator cuff tears," *International Journal of Medical Informatics*, vol. 133, p. 104 015, Jan. 1, 2020. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1386505619304824 (visited on 03/01/2023).

[22] W. Yang and Q. Su, "Process mining for clinical pathway: Literature review and future directions," in *2014 11th International Conference on Service Systems and Service Management (ICSSSM)*, Jun. 2014, pp. 1–5.

[23] K. Monsen, D. Foster, T. Gomez, *et al.*, "Evidence-based Standardized Care Plans for Use Internationally to Improve Home Care Practice and Population Health," *Applied Clinical Informatics*, vol. 2, no. 3, pp. 373–383, Sep. 21, 2011. pmid: 23616884. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3631934/ (visited on 08/31/2023).

[24] Nicholas Oughtibridge. "A system of concepts for the continuity of care," contsys.org. (Oct. 8, 2019), [Online]. Available: https://contsys.org/ (visited on 08/27/2023).

[25] A. P. Kurniati, G. Hall, D. Hogg, and O. Johnson, "Process mining in oncology using the MIMIC-III dataset," *Journal of Physics: Conference Series*, vol. 971, p. 012 008, Mar. 2018. [Online]. Available: https://iopscience.iop.org/article/10.1088/1742-6596/971/1/012008 (visited on 09/14/2022).

[26] R. Lear, L. Freise, M. Kybert, A. Darzi, A. L. Neves, and E. K. Mayer, "Patients' Willingness and Ability to Identify and Respond to Errors in Their Personal Health Records: Mixed Methods Analysis of Cross-sectional Survey Data," *Journal of Medical Internet Research*, vol. 24, no. 7, e37226, Jul. 8, 2022. [Online]. Available: https://www.jmir.org/2022/7/e37226 (visited on 09/21/2022).

[27] B. Goldacre and J. Morley, "Better, Broader, Safer: Using health data for research and analysis. A review commissioned by the Secretary of State for Health and Social Care.," UK Department of Health and Social Care, 2022.

[28] H. Dong, M. Falis, W. Whiteley, *et al.*, "Automated clinical coding: What, why, and where we are?" *NPJ Digital Medicine*, vol. 5, p. 159, Oct. 22, 2022. pmid: 36273236. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC9588058/ (visited on 08/12/2023).

[29] D. G. P. Croft and R. C. of Physicians' iLab, *Engaging clinicians in improving data quality in the NHS*, Centre for Health Information, Research and Evaluation (CHIRAL), Sep. 2006. [Online]. Available: https://www.rcplondon.ac.uk/projects/outputs/engaging-clinicians-improving-data-quality-nhs-ilab-project-summary (visited on 05/28/2022).

[30] S. A. R. Nouraei, J. S. Virk, A. Hudovsky, C. Wathen, A. Darzi, and D. Parsons, "Accuracy of clinician-clinical coder information handover following acute medical admissions: Implication for using administrative datasets in clinical outcomes management," *Journal of Public Health*, vol. 38, no. 2, pp. 352–362, Jun. 2016. [Online]. Available: https://academic.oup.com/jpubhealth/article-lookup/doi/10.1093/pubmed/fdv041 (visited on 04/19/2022).

[31] S. A. Spencer and M. P. Davies, "Hospital episode statistics: Improving the quality and value of hospital data: A national internet e-survey of hospital consultants," *BMJ Open*, vol. 2, no. 6, e001651, 2012. [Online]. Available: https://bmjopen.bmj.com/lookup/doi/10.1136/bmjopen-2012-001651 (visited on 04/19/2022).

[32] Haendel Melissa A., Chute Christopher G., and Robinson Peter N., "Classification, Ontology, and Precision Medicine," *New England Journal of Medicine*, vol. 379, no. 15, pp. 1452–1462, Oct. 11, 2018. [Online]. Available: https://www.nejm.org/doi/full/10.1056/NEJMra1615014 (visited on 07/10/2024).

[33] World Health Organization. "International Statistical Classification of Diseases and Related Health Problems (ICD)," World Health Organization. (2025), [Online]. Available: https://www.who.int/standards/classifications/classification-of-diseases (visited on 06/06/2022).

[34] Terminology and Classifications Delivery Service, *National Clinical Coding Standards ICD-10 (5th Edition)*, NHS Digital, Jun. 2022.

[35] NHS Digital. "Clinical Classifications," NHS Digital. (Jun. 22, 2021), [Online]. Available: https://digital.nhs.uk/services/terminology-and-classifications/clinical-classifications (visited on 06/06/2022).

[36] Terminology and Classifications Delivery Service, "National Clinical Coding Standards OPCS-4 2022: Accurate data for quality information," NHS Digital, 9.1, Jun. 2022.

[37] NHS England. "NHS England - SNOMED CT." (), [Online]. Available: https://www.england.nhs.uk/digitaltechnology/digital-primary-care/snomed-ct/ (visited on 06/06/2022).

[38] J. de Bruin, C. Bos, J. N. Struijs, H. W.-T. Drewes, and C. A. Baan, "Conceptualizing learning health systems: A mapping review," *Learning Health Systems*, vol. n/a, no. n/a, e10311, 2022. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1002/lrh2.10311 (visited on 06/10/2022).

[39] J. Enticott, A. Johnson, and H. Teede, "Learning health systems using data to drive healthcare improvement and impact: A systematic review," *BMC Health Services Research*, vol. 21, no. 1, p. 200, Dec. 2021. [Online]. Available: https://bmchealthservres.biomedcentral.com/articles/10.1186/s12913-021-06215-8 (visited on 08/11/2022).

[40] S. M. Greene, R. J. Reid, and E. B. Larson, "Implementing the Learning Health System: From Concept to Action," *Annals of Internal Medicine*, vol. 157, no. 3, p. 207, Aug. 7, 2012. [Online]. Available: http://annals.org/article.aspx?doi=10.7326/0003-4819-157-3-201208070-00012 (visited on 06/10/2022).

[41] T. G. Erdogan and A. Tarhan, "Systematic Mapping of Process Mining Studies in Healthcare," *IEEE Access*, vol. 6, pp. 24 543–24 567, 2018.

[42] N. Martin, J. De Weerdt, C. Fernández-Llatas, *et al.*, "Recommendations for enhancing the usability and understandability of process mining in healthcare," *Artificial Intelligence in Medicine*, vol. 109, p. 101 962, Sep. 1, 2020. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0933365720312276 (visited on 05/17/2023).

[43] W. M. van der Aalst, "A practitioner's guide to process mining: Limitations of the directly-follows graph," *Procedia Computer Science*, vol. 164, pp. 321–328, 2019. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S1877050919322367 (visited on 12/14/2021).

[44] P. Pijnenborg, R. Verhoeven, M. Firat, H. van Laarhoven, and L. Genga, "Towards Evidence-Based Analysis of Palliative Treatments for Stomach and Esophageal Cancer Patients: A Process Mining Approach," in *2021 3rd International Conference on Process Mining (ICPM)*, Eindhoven, Netherlands: IEEE, Oct. 31, 2021, pp. 136–143. [Online]. Available: https://ieeexplore.ieee.org/document/9576880/ (visited on 12/08/2021).

[45] G. Janssenswillen, B. Depaire, M. Swennen, M. Jans, and K. Vanhoof, "bupaR: Enabling reproducible business process analysis," *Knowledge-Based Systems*, vol. 163, pp. 927–930, Jan. 1, 2019. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0950705118305045 (visited on 01/10/2025).

[46] K. Baker, E. Dunwoodie, R. G. Jones, *et al.*, "Process mining routinely collected electronic health records to define real-life clinical pathways during chemotherapy," *International Journal of Medical Informatics*, vol. 103, pp. 32–41, Jul. 1, 2017. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1386505617300722 (visited on 05/18/2023).

[47] E. Batista and A. Solanas, "Process Mining in Healthcare: A Systematic Review," in *2018 9th International Conference on Information, Intelligence, Systems and Applications (IISA)*, Jul. 2018, pp. 1–6.

[48] A. Berti, S. J. van Zelst, and Wil M.P. van der Aalst. "Process Mining for Python (PM4Py): Bridging the Gap Between Process- and Data Science." arXiv: 1905.06169. (May 15, 2019), pre-published.

[49] A. E. W. Johnson, L. Bulgarelli, L. Shen, *et al.*, "MIMIC-IV, a freely accessible electronic health record dataset," *Scientific Data*, vol. 10, no. 1, p. 1, Jan. 3, 2023. [Online]. Available: https://www.nature.com/articles/s41597-022-01899-x (visited on 10/29/2024).

[50] R. Lenz and M. Reichert, "IT support for healthcare processes – premises, challenges, perspectives," *Data & Knowledge Engineering*, Business Process Management, vol. 61, no. 1, pp. 39–58, Apr. 1, 2007. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0169023X06000784 (visited on 06/01/2023).

[51] L. Chammas, O. Dwyer, E. Sallinger, J. Davies, and E. J. Morris, "Care records and healthcare processes: Adding context to clinical codes," presented at the Hawaii International Conference on System Sciences, 2024. [Online]. Available: http://hdl.handle.net/10125/106829 (visited on 10/30/2024).

[52] Z. Huang, W. Dong, P. Bath, L. Ji, and H. Duan, "On mining latent treatment patterns from electronic medical records," *Data Mining and Knowledge Discovery*, vol. 29, no. 4, pp. 914–949, Jul. 1, 2015. [Online]. Available: https://doi.org/10.1007/s10618-014-0381-y (visited on 03/01/2023).

[53] H. Yan, P. Van Gorp, U. Kaymak, *et al.*, "Aligning Event Logs to Task-Time Matrix Clinical Pathways in BPMN for Variance Analysis," *IEEE Journal of Biomedical and Health Informatics*, vol. 22, no. 2, pp. 311–317, Mar. 2018. [Online]. Available: http://ieeexplore.ieee.org/document/8039478/ (visited on 11/17/2021).

[54] The University of Texas MD Anderson Cancer Center, *Colon Cancer Treatment Clinical Algorithms*, The University of Texas MD Anderson Cancer Center, Nov. 15, 2022. [Online]. Available: https://www.mdanderson.org/content/dam/mdanderson/documents/for-physicians/algorithms/cancer-treatment/ca-treatment-colon-web-algorithm.pdf (visited on 01/02/2025).

[55] T. Tamrat, N. Ratanaprayul, M. Barreix, *et al.*, "Transitioning to Digital Systems: The Role of World Health Organization's Digital Adaptation Kits in Operationalizing Recommendations and Interoperability Standards," *Global Health: Science and Practice*, vol. 10, no. 1, Feb. 28, 2022. pmid: 35294382. [Online]. Available: https://www.ghspjournal.org/content/10/1/e2100320 (visited on 12/14/2024).

[56] M. Michaels, "Adapting Clinical Guidelines for the Digital Age: Summary of a Holistic and Multidisciplinary Approach," *American Journal of Medical Quality*, vol. 38, no. 5, S3–S11, 2023. pmid: 37668270. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC10476590/ (visited on 12/14/2024).

[57] Isabel dos Santos Silva, *Cancer Epidemiology: Principles and Methods*. Lyon: International Agency for Research on Cancer, 1999, 442 pp.

[58] D. J. Spiegelhalter, "Funnel plots for comparing institutional performance," *Statistics in Medicine*, vol. 24, no. 8, pp. 1185–1202, Apr. 30, 2005. [Online]. Available: https://onlinelibrary.wiley.com/doi/10.1002/sim.1970 (visited on 12/23/2024).

[59] J. T. Rich, J. G. Neely, R. C. Paniello, C. C. J. Voelker, B. Nussenbaum, and E. W. Wang, "A practical guide to understanding Kaplan-Meier curves.," *Otolaryngology–head and neck surgery : official journal of American Academy of Otolaryngology-Head and Neck Surgery*, vol. 143, no. 3, pp. 331–336, Sep. 2010. pmid: 20723767.

[60] Cancer Research UK. "Cancer Statistics for the UK," Cancer Research UK. (Dec. 9, 2024), [Online]. Available: https://www.cancerresearchuk.org/health-professional/cancer-statistics-for-the-uk (visited on 12/09/2024).

[61] NHS Cancer Programme, *Implementing a timed colorectal cancer diagnostic pathway: A handbook for local health and care systems*, NHS England, Apr. 2018. [Online]. Available: https://www.england.nhs.uk/wp-content/uploads/2018/04/implementing-timed-colorectal-cancer-diagnostic-pathway.pdf.

[62] Cancer Research UK (CRUK). "Cancer waiting times." (Dec. 9, 2024), [Online]. Available: https://www.cancerresearchuk.org/about-cancer/worried-about-cancer/cancer-waiting-times (visited on 11/28/2024).

[63] The British Medical Association. "NHS backlog data analysis," The British Medical Association. (Dec. 16, 2024), [Online]. Available: https://www.bma.org.uk/advice-and-support/nhs-delivery-and-workforce/pressures/nhs-backlog-data-analysis (visited on 01/09/2025).

[64] Bowel Cancer UK, *Delivering the ambitions of the NHS Long Term Plan: A review of progress on bowel cancer in England*, Dec. 2024.

[65] National Institute for Health and Care Excellence. "Colorectal cancer NICE Guidance [NG151]." (Dec. 15, 2021), [Online]. Available: https://www.nice.org.uk/guidance/ng151 (visited on 01/29/2024).

[66] Chelsea and Westminster Hospital NHS Foundation Trust. "Role of the colorectal multidisciplinary team," Chelsea and Westminster Hospital NHS Foundation Trust. (2024), [Online]. Available: https://www.chelwest.nhs.uk/your-visit/patient-leaflets/surgery-services/role-of-the-colorectal-multidisciplinary-team (visited on 12/09/2024).

[67] NHS England. "Cancer Alliances – improving care locally," NHS England. (), [Online]. Available: https://www.england.nhs.uk/cancer/cancer-alliances-improving-care-locally/ (visited on 01/10/2025).

[68] K. E. Henson, L. Elliss-Brookes, V. H. Coupland, *et al.*, "Data Resource Profile: National Cancer Registration Dataset in England," *International Journal of Epidemiology*, vol. 49, no. 1, 16–16h, Feb. 1, 2020. [Online]. Available: https://academic.oup.com/ije/article/49/1/16/5476570 (visited on 04/19/2022).

[69] C. J. Bright, S. Lawton, S. Benson, *et al.*, "Data Resource Profile: The Systemic Anti-Cancer Therapy (SACT) dataset," *International Journal of Epidemiology*, vol. 49, no. 1, pp. 15–15l, Feb. 1, 2020. [Online]. Available: https://academic.oup.com/ije/article/49/1/15/5538002 (visited on 04/14/2022).

[70] S. Sandhu, M. Sharpe, Ú. Findlay, *et al.*, "Cohort profile: Radiotherapy dataset (RTDS) in England," *BMJ Open*, vol. 13, no. 6, e070699, Jun. 1, 2023. pmid: 37339842. [Online]. Available: https://bmjopen.bmj.com/content/13/6/e070699 (visited on 01/10/2025).

[71] A. Boyd, R. Cornish, L. Johnson, *et al.*, "Understanding Hospital Episode Statistics (HES)," CLOSER, 2017. [Online]. Available: https://www.closer.ac.uk/wp-content/uploads/%20CLOSER-resource-understanding-hospital-episode-statistics-2018.pdf.

[72] A. Downing, P. Hall, R. Birch, *et al.*, "Data Resource Profile: The COloRECTal cancer data repository (CORECT-R)," *International Journal of Epidemiology*, vol. 50, no. 5, 1418–1418k, Nov. 10, 2021. pmid: 34255059.

[73] Thames Valley and Surrey. "Thames Valley and Surrey Secure Data Environment (SDE)." (), [Online]. Available: https://thamesvalleyandsurreyhealthandcaredata.nhs.uk/using-patient-data/thames-valley-and-surrey-secure-data-environment/ (visited on 12/30/2024).

[74] S. Remy, L. Pufahl, J. P. Sachs, E. Böttinger, and M. Weske, "Event log generation in a health system: A case study," in *Business Process Management: 18th International Conference, BPM 2020, Seville, Spain, September 13–18, 2020, Proceedings 18*, Springer, 2020, pp. 505–522.

[75] U. Kaymak, R. Mans, T. van de Steeg, and M. Dierks, "On process mining in health care," in *2012 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, Oct. 2012, pp. 1859–1864.

[76] J. Cremerius, L. Pufahl, F. Klessascheck, and M. Weske, "Event log generation in MIMIC-IV research paper," in *Process Mining Workshops (2022)*, ser. Lecture Notes in Business Information Processing, vol. 468, Springer, 2022, pp. 302–314.

[77] B.-H. Chew, "Planning and Conducting Clinical Research: The Whole Process," *Cureus*, Feb. 20, 2019. [Online]. Available: https://www.cureus.com/articles/17094-planning-and-conducting-clinical-research-the-whole-process (visited on 08/23/2023).

[78] J. Cremerius, H. Patzlaff, V. Rahn, H. Leopold, and K. Logistics, "Data-Based Process Variant Analysis," in *Proceedings of the 56th Hawaii International Conference on System Sciences*, 2023. [Online]. Available: https://hdl.handle.net/10125/103031%20978-0-9981331-6-4.

[79] A. Herbert, L. Wijlaars, A. Zylbersztejn, D. Cromwell, and P. Hardelid, "Data Resource Profile: Hospital Episode Statistics Admitted Patient Care (HES APC)," *International Journal of Epidemiology*, vol. 46, no. 4, pp. 1093–1093i, Aug. 1, 2017. [Online]. Available: https://academic.oup.com/ije/article/46/4/1093/3072145 (visited on 04/14/2022).

[80] J. C. Taylor, D. Swinson, J. F. Seligmann, *et al.*, "Addressing the variation in adjuvant chemotherapy treatment for colorectal cancer: Can a regional intervention promote national change?" *International Journal of Cancer*, vol. 148, no. 4, pp. 845–856, Feb. 15, 2021. [Online]. Available: https://onlinelibrary.wiley.com/doi/10.1002/ijc.33261 (visited on 10/26/2021).

[81] E. Aspland, P. R. Harper, D. Gartner, P. Webb, and P. Barrett-Lee, "Modified Needleman–Wunsch algorithm for clinical pathway clustering," *Journal of Biomedical Informatics*, vol. 115, p. 103 668, Mar. 1, 2021. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1532046420302963 (visited on 02/20/2023).

[82] E. J. A. Morris, P. J. Finan, K. Spencer, *et al.*, "Wide Variation in the Use of Radiotherapy in the Management of Surgically Treated Rectal Cancer Across the English National Health Service," *Clinical Oncology*, vol. 28, no. 8, pp. 522–531, Aug. 1, 2016. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0936655516000662 (visited on 03/29/2023).

[83] J. C. Taylor, L. H. Iversen, D. Burke, *et al.*, "Influence of age on surgical treatment and postoperative outcomes of patients with colorectal cancer in Denmark and Yorkshire, England," *Colorectal Disease*, vol. 23, no. 12, pp. 3152–3161, 2021. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1111/codi.15910 (visited on 04/24/2023).

[84] E. Syriopoulou, E. Morris, P. J. Finan, P. C. Lambert, and M. J. Rutherford, "Understanding the impact of socioeconomic differences in colorectal cancer survival: Potential gain in life-years," *British Journal of Cancer*, vol. 120, no. 11, pp. 1052–1058, May 2019. [Online]. Available: http://www.nature.com/articles/s41416-019-0455-0 (visited on 04/24/2023).

[85] L. Chammas. "Phd and Publication Codelists," GitHub. (Sep. 28, 2023), [Online]. Available: https://github.com/larachammas/Phd-and-Publication-Codelists (visited on 10/29/2024).

[86] National Disease Registration Service (NDRS). "Cancer Treatments," Treatment Breakdown by Demographics, 2013-2021. (), [Online]. Available: https://nhsd-ndrs.shinyapps.io/treatment_by_demographic_factors/ (visited on 12/30/2024).

[87] A. Shrestha, C. Martin, M. Burton, S. Walters, K. Collins, and L. Wyld, "Quality of life versus length of life considerations in cancer patients: A systematic literature review," *Psycho-Oncology*, vol. 28, no. 7, pp. 1367–1380, Jul. 2019. pmid: 30838697. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6619389/ (visited on 06/14/2023).

[88] A. Gasparini, "Comorbidity: An R package for computing comorbidity scores," *Journal of Open Source Software*, vol. 3, no. 23, p. 648, 2018. [Online]. Available: https://doi.org/10.21105/joss.00648.

[89] M. E. Charlson, P. Pompei, K. L. Ales, and C. R. MacKenzie, "A new method of classifying prognostic comorbidity in longitudinal studies: Development and validation," *Journal of Chronic Diseases*, vol. 40, no. 5, pp. 373–383, 1987. pmid: 3558716.

[90] "Chemotherapy Regimens Clinical Coding Standards and Guidance OPCS-4," Health and Social Care Information Centre, Apr. 2017.

[91] T. Grisold, H. van der Aa, S. Franzoi, S. Hartl, J. Mendling, and J. V. Brocke, "A Context Framework for Sense-making of Process Mining Results," in *2024 6th International Conference on Process Mining (ICPM)*, Oct. 2024, pp. 57–64. [Online]. Available: https://ieeexplore.ieee.org/document/10680665/authors#authors (visited on 01/10/2025).

[92] NHS Digital, *SNOMED CT to ICD-10 and OPCS-4 Map Table Technical Specification and Implementation Guidance (UK Edition)*, Jul. 2021.

[93] Observational Health Data Sciences and Informatics. "Standardized Vocabularies," GitHub. (Dec. 2024), [Online]. Available: https://github.com/OHDSI/Vocabulary-v5.0/wiki/Standardized-Vocabularies (visited on 11/06/2024).

[94] O. P. Dwyer, L. Chammas, E. Sallinger, and J. Davies, "Investigating an Ontology-Informed Approach to Event Log Generation in Healthcare," in *Process Mining Workshops*, J. De Smedt and P. Soffer, Eds., Cham: Springer Nature Switzerland, 2024, pp. 235–246.

[95] H. De Oliveira, M. Prodel, L. Lamarsalle, *et al.*, ""Bow-tie" optimal pathway discovery analysis of sepsis hospital admissions using the Hospital Episode Statistics database in England," *JAMIA open*, vol. 3, no. 3, pp. 439–448, Oct. 2020. pmid: 33215077.

[96] J. J. Koorn, I. Beerepoot, V. S. Dani, *et al.*, "Bringing Rigor to the Qualitative Evaluation of Process Mining Findings: An Analysis and a Proposal," in *2021 3rd International Conference on Process Mining (ICPM)*, Eindhoven, Netherlands: IEEE, Oct. 31, 2021, pp. 120–127. [Online]. Available: https://ieeexplore.ieee.org/document/9576877/ (visited on 12/08/2021).

[97] R. J. C. Bose, R. S. Mans, and W. M. van der Aalst, "Wanna improve process mining results?" In *2013 IEEE Symposium on Computational Intelligence and Data Mining (CIDM)*, Apr. 2013, pp. 127–134.

[98] H. M. Marin-Castro and E. Tello-Leal, "Event Log Preprocessing for Process Mining: A Review," *Applied Sciences*, vol. 11, no. 22, p. 10 556, 22 Jan. 2021. [Online]. Available: https://www.mdpi.com/2076-3417/11/22/10556 (visited on 10/29/2024).

[99] A. K. Alves de Medeiros and W. M. P. van der Aalst, "Process Mining towards Semantics," in *Advances in Web Semantics I: Ontologies, Web Services and Applied Semantic Web*, T. S. Dillon, E. Chang, R. Meersman, and K. Sycara, Eds., Berlin, Heidelberg: Springer, 2009, pp. 35–80. [Online]. Available: https://doi.org/10.1007/978-3-540-89784-2_3 (visited on 10/29/2024).

[100] G. Leonardi, M. Striani, S. Quaglini, A. Cavallini, and S. Montani, "Towards Semantic Process Mining Through Knowledge-Based Trace Abstraction," in *Data-Driven Process Discovery and Analysis*, P. Ceravolo, M. van Keulen, and K. Stoffel, Eds., Cham: Springer International Publishing, 2019, pp. 45–64.

[101] F. Klessascheck, T. Lichtenstein, M. Meier, *et al.*, "Domain-specific event abstraction," in *Business Information Systems*, 2021, pp. 117–126.

[102] J. Watson, B. D. Nicholson, W. Hamilton, and S. Price, "Identifying clinical features in primary care electronic health record studies: Methods for codelist development," *BMJ Open*, vol. 7, no. 11, e019637, Nov. 1, 2017. pmid: 29170293. [Online]. Available: https://bmjopen.bmj.com/content/7/11/e019637 (visited on 08/25/2023).

[103] M. Elkheder, A. Gonzalez-Izquierdo, M. Qummer Ul Arfeen, *et al.*, "Translating and evaluating historic phenotyping algorithms using SNOMED CT," *Journal of the American Medical Informatics Association*, vol. 30, no. 2, pp. 222–232, Feb. 1, 2023. [Online]. Available: https://doi.org/10.1093/jamia/ocac158 (visited on 08/12/2023).

[104] NHS Digital, *NHS Digital Terminology Server Github*, NHS Digital, May 1, 2024. [Online]. Available: https://github.com/NHSDigital/TerminologyServer (visited on 12/11/2024).

[105] *Expression Constraint Language - Specification and Guide - SNOMED CT*, International Health Terminology Standards Development Organisation, Nov. 23, 2023. [Online]. Available: https://confluence.ihtsdotools.org/display/DOCECL/Expression+Constraint+Language+-+Specification+and+Guide (visited on 12/11/2024).

[106] Health Data Research UK. "HDRUK Phenotype Library." (), [Online]. Available: http://phenotypes.healthdatagateway.org/ (visited on 01/03/2025).

[107] C. Reich, A. Ostropolets, P. Ryan, *et al.*, "OHDSI Standardized Vocabularies—a large-scale centralized reference ontology for international data harmonization," *Journal of the American Medical Informatics Association*, vol. 31, no. 3, pp. 583–590, Mar. 1, 2024. [Online]. Available: https://doi.org/10.1093/jamia/ocad247 (visited on 01/07/2025).

[108] M. I. M. Ilyas, "Epidemiology of Stage IV Colorectal Cancer: Trends in the Incidence, Prevalence, Age Distribution, and Impact on Life Span," *Clinics in Colon and Rectal Surgery*, vol. 37, no. 2, pp. 57–61, Mar. 15, 2023. pmid: 38322602. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC10843881/ (visited on 12/23/2024).

[109] H. M. Fenton, P. J. Finan, R. Milton, *et al.*, "National variation in pulmonary metastasectomy for colorectal cancer," *Colorectal Disease*, vol. 23, no. 6, pp. 1306–1316, Jun. 2021. [Online]. Available: https://onlinelibrary.wiley.com/doi/10.1111/codi.15506 (visited on 03/13/2023).

[110] M. Araghi, M. Arnold, M. J. Rutherford, *et al.*, "Colon and rectal cancer survival in seven high-income countries 2010–2014: Variation by age and stage at diagnosis (the ICBP SURVMARK-2 project)," *Gut*, vol. 70, no. 1, pp. 114–126, Jan. 1, 2021. pmid: 32482683. [Online]. Available: https://gut.bmj.com/content/70/1/114 (visited on 12/23/2024).

[111] O. Hernandez Dominguez, S. Yilmaz, and S. R. Steele, "Stage IV Colorectal Cancer Management and Treatment," *Journal of Clinical Medicine*, vol. 12, no. 5, p. 2072, 5 Jan. 2023. [Online]. Available: https://www.mdpi.com/2077-0383/12/5/2072 (visited on 07/16/2024).

[112] A. Downing, A. W. Glaser, P. J. Finan, *et al.*, "Functional Outcomes and Health-Related Quality of Life After Curative Treatment for Rectal Cancer: A Population-Level Study in England," *International Journal of Radiation Oncology, Biology, Physics*, vol. 103, no. 5, pp. 1132–1142, Apr. 2019. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S0360301618341221 (visited on 03/29/2023).

[113] N. C. Buchs, F. Ris, P. E. Majno, *et al.*, "Rectal Outcomes After a Liver-First Treatment of Patients with Stage IV Rectal Cancer," *Annals of Surgical Oncology*, vol. 22, no. 3, pp. 931–937, Mar. 2015. [Online]. Available: http://link.springer.com/10.1245/s10434-014-4069-8 (visited on 12/23/2024).

[114] D. Harji, A. Vallance, T. Ibitoye, *et al.*, "IMPACT organizational survey highlighting provision of services for patients with locally advanced and recurrent colorectal cancer across Great Britain and Ireland," *Colorectal Disease*, vol. 26, no. 12, pp. 2033–2038, 2024. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1111/codi.17139 (visited on 12/26/2024).

[115] H. M. Fenton, J. C. Taylor, J. P. A. Lodge, *et al.*, "Variation in the Use of Resection for Colorectal Cancer Liver Metastases," *Annals of Surgery*, vol. 270, no. 5, pp. 892–898, Nov. 2019. [Online]. Available: https://journals.lww.com/10.1097/SLA.0000000000003534 (visited on 03/13/2023).

[116] A. E. Vallance, J. van der Meulen, A. Kuryba, *et al.*, "Socioeconomic differences in selection for liver resection in metastatic colorectal cancer and the impact on survival," *European Journal of Surgical Oncology: The Journal of the European Society of Surgical Oncology and the British Association of Surgical Oncology*, vol. 44, no. 10, pp. 1588–1594, Oct. 2018. pmid: 29895508.

[117] S. M. Crawford, V. Sauerzapf, R. Haynes, D. Forman, and A. P. Jones, "Social and geographical factors affecting access to treatment of colorectal cancer: A cancer registry study," *BMJ Open*, vol. 2, no. 2, e000410, 2012. [Online]. Available: https://bmjopen.bmj.com/lookup/doi/10.1136/bmjopen-2011-000410 (visited on 12/10/2024).

[118] K. Yadav and R. J. Lewis, "Immortal Time Bias in Observational Studies," *JAMA*, vol. 325, no. 7, pp. 686–687, Feb. 16, 2021. [Online]. Available: https://doi.org/10.1001/jama.2020.9151 (visited on 12/23/2024).

[119] C. Fernandez-Llatas, B. Valdivieso, V. Traver, and J. M. Benedi, "Using Process Mining for Automatic Support of Clinical Pathways Design," in *Data Mining in Clinical Medicine*, C. Fernández-Llatas and J. M. García-Gómez, Eds., vol. 1246, New York, NY: Springer New York, 2015, pp. 79–88. [Online]. Available: http://link.springer.com/10.1007/978-1-4939-1985-7_5 (visited on 12/06/2021).

[120] N. Iglesias, J. M. Juarez, and M. Campos, "Business Process Model and Notation and openEHR Task Planning for Clinical Pathway Standards in Infections: Critical Analysis," *Journal of Medical Internet Research*, vol. 24, no. 9, e29927, Sep. 15, 2022. [Online]. Available: https://www.jmir.org/2022/9/e29927 (visited on 03/01/2023).