Methods for analysing care pathways

Ontology, representation, and process perspectives on health data

Owen Patrick Dwyer

GREEN TEMPLETON COLLEGE
UNIVERSITY OF OXFORD

A thesis submitted for the degree
of Doctor of Philosophy

Michaelmas 2024

Abstract

Every interaction with the healthcare system leaves some kind of trace in the form of data, which forms a valuable resource for research. Modern healthcare is about more than individual interventions and diagnoses: equally important is the combination, ordering and timing of events, and the way that the patient moves through the system: the *patient pathway*. This is the source of a great many research questions, which are difficult to answer: patients cannot be neatly divided into groups, and "compliance" with a particular standard is hard to measure when there are many decision points and unseen variables.

I first examine the extent to which contextual factors affect treatment decisions, and demonstrate that pathway data is shaped by human practices, processes and biases. I then consider the logic behind which data is included in process analysis, and propose an approach that uses ontological knowledge to infer relationships between diagnoses and procedures.

I then propose *embedding-based dynamic time warping* (E-DTW), an algorithm for describing the similarity between two patients' pathways. This algorithm is designed with practical characteristics in mind: it incorporates information on both the semantic similarity of the events in a pathway and their temporal patterns, it uses knowledge from standard and publicly available ontologies, and its embeddings can be re-used for different tasks.

Finally, I extend the E-DTW method to measure the semantic similarity between a patient's pathway and a pathway as laid down in guidelines; in the process, I describe a set of steps for assessing the gap between a guideline and given dataset, and a notation for encoding pathway guidelines in computable form.

Structured data is, by virtue of the way it is recorded and encoded, rich in semantics that can be exploited to create useful insights. Pathways are inherent complex, and purely logical or statistical attempts to analyse them have drawbacks. This thesis combines the use of modern and flexible machine learning concepts with grounding in ontological knowledge, describing a set of methods that allow the benefits of health data to be realised whilst also ensuring that analysis is relevant, reliable, and reproducible.

Acknowledgements

> *I don't know half of you half as well as I should like; and I like less than half of you half as well as you deserve.*
>
> J.R.R. Tolkien, *The Fellowship of the Ring*

This thesis would not have been possible without the support and guidance of my supervisors, Professors Emanuel Sallinger and Jim Davies, over the last four years. Thank you for everything you have taught me, which include — in no particular order — that deadlines are mostly just guidelines, that it's all just nodes and edges, and that there is no such thing as a free lunch. I'm not sure that they were the lessons I was supposed to get out of this, but I'm grateful nonetheless.

I am also indebted to Eva Morris, Max Van Kleek, and Niels Martin, whose comments and feedback were extremely valuable. I also need to acknowledge my previous academic supervisors at Lancaster for getting me this far: Jo Knight, Peter Diggle, and in particular Angelos Marnerides, for encouraging me to do a PhD in the first place, and for telling me to write the paper that started all of this.

I'm also grateful to Lara Chammas for being an excellent co-author, conference buddy, and proofreader, and for her endless positivity. I'd also like to thank Adam Sturge for all the lunches. Perhaps one day we'll find that seafood macaroni again.

This thesis would not have been possible without financial support from Elsevier and the Engineering and Physical Sciences Research Council. I also appreciate the support of Elia Lima-Walton and her team at Elsevier, and the entire NIHR Health Informatics Collaborative and the Thames Valley and Surrey Secure Data Environment teams. You're stuck with me now. Sorry.

My deepest gratitude goes to my family and friends, for their endless support and encouragement which has kept me (mostly) sane during these four years. Finally, to Hannah, who has been there every step of the way. Thank you for holding my hand.

Contents

List of figures

List of tables

List of abbreviations

2WW two-week wait
ACPGBI Association of Coloproctology of Great Britain and Ireland
AI artificial intelligence
BPMN Business Process Model and Notation
CIG computer-interpretable guideline
CORECT-R Colorectal Cancer Data Repository
CRT chemoradiotherapy
DHSC Department of Health and Social Care
DFG directly-follows graph
DTW dynamic time warping
DL Damerau-Levenshtein distance
ECL (SNOMED CT) Expression Constraint Language
E-DTW embedding-based dynamic time warping
EHR electronic health record
ESMO European Society for Medical Oncology
FIT faecal immunochemical test
GP general practitioner
HIC (NIHR) Health Informatics Collaborative
ICD International Classification of Diseases
IMD Index of Multiple Deprivation
KG knowledge graph
KGE knowledge graph embedding
LCS longest common subsequence
LLM large language model
MCL Markov cluster algorithm
MDT multidisciplinary team
MGD minimum guideline distance
ML machine learning
MRR mean reciprocal rank
NHS National Health Service
NICE National Institute for Health and Care Excellence
NIHR National Institute for Health and Care Research
NLP natural language processing
NW Needleman-Wunsch algorithm

OMOP CDM Observational Medical Outcomes Partnership Common Data Model

OPCS Office of Population Censuses and Surveys (Classification of Interventions and Procedures)

OUH Oxford University Hospitals NHS Foundation Trust

PCA principle component analysis

PM process mining

PWF Pseudo-Workflow

RBO rank-biased overlap

SCPRT short-course preoperative radiotherapy

SNOMED CT Systematized Nomenclature of Medicine Clinical Terms

SSA state sequence analysis

UML Unified Modelling Language

UMLS Unified Medical Language System

1 *Introduction*

The healthcare system in the United Kingdom collects massive volumes of data every day; almost every interaction with the health service leaves some kind of trace. In recent years, the idea of reusing this routinely collected data for research, analysis, and service improvement has become a significant area of focus. Such data has enormous potential: it can be used to assess the efficacy of medications and interventions, it can be used to monitor the quality of care being provided, and it can be used to measure the impact of particular policies and practices, to name but a few examples.

However, modern healthcare is about more than individual interventions and diagnoses. Attention also needs to be paid to the entire journeys of patients: the ways in which they move through the system, the combinations and ordering of events that occur in their journey, and the time they take to get there. These elements, taken together, are said to form the *patient pathway*. Many research questions naturally surround patient pathways: for example, how many patients follow a given pathway, which factors influence the choice, timing, and order of treatment, and whether particular patterns are associated with particular outcomes. Many healthcare providers have, explicitly or implicitly, some concept of the recommended route through treatment, which serves to guides decision-making and provides a benchmark for best — or at least usual — practice. The existence of such guidelines prompts further questions, in particular regarding whether compliance with a particular pathway policy actually improves outcomes.

These questions are difficult to evaluate for a number of reasons. Data is often incomplete, or collected for different purposes or in a different form than what is needed to answer these questions. Where data is collected, the volume is often overwhelming: given thousands of patients' complete records, researchers have to make judgements around which sorts of events are included in analysis, and which are not. Methods for analysis are often simplistic: dividing patients into "pathway" and "non-pathway" cohorts, or excluding and including discrete events without considering subjective and semantic factors such as conceptual similarity lead to the oversimplification of complex processes, and the loss of valuable domain knowledge.

It is this set of challenges that this thesis aims to address. By drawing on methods for representing semantic meaning and context, particularly ontologies and knowledge graphs, and by applying them to real datasets, guidelines, and research questions, I aim to design a set of analysis techniques that address the shortcomings of existing methods and support the effective analysis of healthcare data from a pathway-oriented perspective.

1.1 *Context*

Throughout this thesis, I use colorectal cancer as a motivating example and as a source of case studies. Colorectal cancer — also commonly referred to as bowel cancer — is the fourth most commonly diagnosed cancer in the UK, with over 44,000 new cases diagnosed every year (Cancer Research UK 2015). Whilst the vast majority of cases occur in those aged 50 and above, there is evidence that incidence is increasing in younger populations (Sung et al. 2024). The term encompasses both colon and rectal cancer: whilst they are two distinct sites with differences in treatment, they neighbour each other, share common screening techniques, and are often grouped together for statistical and research purposes.

Colorectal cancer is the second most common cause of cancer death in the UK, and survival rates are lower in the UK than in comparable countries (Allemani et al. 2018; Coleman et al. 2011). There is significant heterogeneity in approaches to treatment, including differences observed in chemotherapy (Boyle et al. 2020; Taylor et al. 2021), radiotherapy (E. J. A. Morris et al. 2016), and surgery (Fenton et al. 2021; E. J. A. Morris et al. 2008), meaning that there are many outstanding research questions regarding the optimal treatment pathway.

1.2 *Contributions*

The main aim of this thesis is to develop a set of methods that support and enhance the analysis of retrospective data from electronic health records (EHRs), and in particular that enable analysis from a pathway perspective. This thesis makes a number of contributions to research in process-oriented health data science:

A practical understanding of pathways — which questions, and which data?

Research studies have inconsistent definitions of what constitutes a pathway, what data is required to analyse them, and which subset of events

in a patient's history are relevant to analysis. I present an argument and evidence that pure event data is rarely either objective or enough: such data is shaped by human practices, processes, and biases, and therefore requires an understanding of its context to be interpreted in a useful way. I also describe a standardised method for determining, with the support of codified domain knowledge from ontologies, which data points to include and exclude — a vital part of analysis often inconsistently performed and documented. These two studies demonstrate how process-oriented data research can more closely align itself to real research questions in healthcare and epidemiology, and how shared definitions of pathways and their contents can make analysis more transparent and reproducible.

An evaluation of the suitability of knowledge graph embedding algorithms for creating meaningful representations of healthcare concepts

A drawback of current methods for measuring pathway similarity is their tendency to create binary divisions, for example between patients "following" or "not following" a pathway, or between interventions that are either "acceptable" or "not acceptable" components of the pathway. This ignores the fact that similarity and relevance are not binary, and that some concepts can be inherently more or less similar to others. For this reason, vector embeddings that describe concept similarity are potentially a very useful tool; I therefore examine KGEs, a set of embedding algorithms that focus on graph-encoded knowledge, and investigate how effective they are at creating meaningful representations of concepts from SNOMED CT using a rigorous evaluation approach that combines two existing evaluation frameworks.

A method for measuring similarity between two patient pathways

Continuing the idea that semantic similarity should be an important part of pathway methods, I propose *embedding-based dynamic time warping* (E-DTW), a measure of similarity between patient pathways that takes into account both the semantic similarity of the events involved and the timing between them. This approach is designed with a number of features in mind to make it practical and advantageous for deployment in real healthcare systems.

An analysis of the relationship between care pathways and retrospective datasets

As well as examining how similar patients' pathways are to each other, many research questions involve measuring how similar real pathways are to recommended pathways. Previous formalisms for representing guide-

lines typically focus on individual intervention points and decision support applications, and have not been widely adopted. I propose a language for representing pathway guidelines that is specifically tailored to the requirements of whole-pathway analysis in large retrospective datasets: it encodes temporal constraints that are important to pathways, and it supports the use of the E-DTW metric, allowing patient populations to be compared against pathways at scale.

1.3 *Publications*

Elements of this thesis have previously been published in peer-reviewed papers. Section 3.2 is based on work presented at HICSS 2024 as *Care records and healthcare processes: adding context to clinical codes* (Chammas et al. 2024). Lara Chammas was responsible for conceiving the original methodology and principles, and leading the writing of the paper including the interpretation of results; I was responsible for all data preparation, coding and analysis, as well as contributions to writing. Emanuel Sallinger, Jim Davies, and Eva Morris provided overall supervision and feedback.

Section 3.3 is based on work presented at PODS4H 2023 as *Investigating an ontology-informed approach to event log generation in healthcare* (Dwyer et al. 2024). I conceived the original method, prepared all data and code, and performed all analysis. Lara Chammas provided assistance in writing and conceptualisation; Emanuel Sallinger and Jim Davies provided overall supervision and feedback. The version described in this thesis significantly develops this work, expanding the scope of the method, and evaluates it in a greater number of scenarios. Since the original submission of this thesis, this follow-up has been published in the Journal of Intelligent Information Systems as *Using ontologies to facilitate healthcare process mining and analysis* (Dwyer et al. 2025).

I also led the writing of an additional paper, *Reasoning over health records with Vadalog: a rule-based approach to patient pathways* (Dwyer et al. 2023), which won the RULEML+RR 2023 Rule Challenge. This work is not detailed in this paper, since it largely deals with methods outside of the scope of this thesis; it is, however, cited as supporting evidence in Sections 2.5 and 5.1.3.

1.4 *Structure*

The relevant background material that this thesis builds on is introduced in Chapter 2. It introduces the concept of clinical pathways and the ways in which they are typically evaluated. It then provides an overview of relevant

work on computationally modelling and comparing pathways, including process mining and machine learning methods. It then introduces knowledge graph methods and their applications in health data, as well as considering the practical factors in implementing AI and data solutions in healthcare systems.

Chapter 3 presents two different views on the definition and construction of patient pathways. It discusses pathway data from the point of view of real implementation in the healthcare system, including the ways in which coded and structured clinical data is influenced by contextual and human factors, and examines the extent to which these factors affect real process models. Firstly, pathways are examined from a data-oriented perspective, and consider which data points should be analysed to create meaningful models. Secondly a method is proposed for automatically preparing process models from raw datasets with the help of domain knowledge from formal ontologies, and is demonstrated in practice.

Chapter 4 proposes E-DTW, a method for quantifying the similarity between patient pathways. It compares several existing methods for creating low-dimensional representations of concepts, focusing on knowledge graph embeddings. Several KGE approaches are compared and evaluated based on their ability to represent concepts from the SNOMED CT ontology. Appendix A includes detailed results each of the KGE approaches evaluated, including a breakdown by individual code chapters, and comparisons of within- and between group distances. A method is then outlined which composes these representations into pathways, and measures the differences between them. This method is then applied to a dataset of rectal cancer pathways; several different variations of the model are tested and compared, and an extended comparison is included in Appendix B. The method is used to discover natural clusters in the data, which are examined to identify how differences in patient characteristics related to differences in treatment, and are compared to known phenotypes from previous research on the subject.

Chapter 5 extends this method to the related problem of measuring similarity between observed pathways in data and idealised pathways described in guidelines. It examines existing methods for encoding guidelines, considering their advantages and disadvantages. Following this, it considers the specific characteristics of clinical guidelines in the UK and the needs of pathway-focused research, and describes a notation for representing pathways tailored to these needs. A method is then proposed for converting these representations into low-dimensional form, such that they can be compared using the E-DTW measure; this is again applied to a set of rectal cancer pathways and the research implications are discussed.

Finally, Chapter 6 discusses this thesis' key findings, their relationship to each other, their contributions to the field, and their implications for researchers and practitioners. I identify five key messages from the experiments described here, as well as several promising avenues for future research.

2 *Literature review*

This chapter provides background information on a number of relevant areas of research to this thesis. It firstly introduces the broad context of data-driven healthcare research in the UK, and the practical considerations and limitations involved in implementing data-intensive methods in a healthcare system. It then introduces the concept of clinical pathways in detail, by examining the existing literature on their definition and implementation, and particularly focuses on the current methods that are used to analyse and evaluate them including both their advantages and drawbacks. It then moves on to discuss several methodologies that offer potential solutions, in particular the various methods that have been proposed for the analysis of either pathways or similarly structured problems such as disease trajectories or sequences of states. Finally, key concepts surrounding knowledge graphs and the Unified Modelling Language standard are introduced.

2.1 *Data-driven research in the NHS*

In recent years, significant attention has been paid to the vast amounts of data created and stored by the UK's National Health Service (NHS), and much has been made of its potential to support a variety of endeavours: in analysing and planning service provision, in identifying patients who would benefit from new interventions or screening programmes, in identifying participants in clinical trials, and in conducting large-scale population health studies (Sudlow 2024). In parallel, the recent boom in artificial intelligence (AI) has led to significant interest in the use of AI to support medical research and treatment, with the British government pledging millions for AI-focused projects (Department of Health and Social Care (DHSC) 2023).

Despite these opportunities, there are many barriers to the use of health data for research. Some of these are structural: the NHS's data infrastructure is highly fragmented, consisting of 7,000 individual data controllers, which makes gathering data on a scale sufficient to perform population-level analyses or to train advanced models very difficult (Zhang et al. 2023). There are also limitations of the data itself. Developers of AI models often assume that perfect and complete data exists in place — in reality, properly structured

and coded data often doesn't exist, or not at the necessary scale (Bainbridge 2019).

The data that exists is also not a perfect and complete record of the decision making process. EHR systems are not designed for research first — they are designed as a note-taking tool for clinicians (Goldacre & J. Morley 2022). Events can go unrecorded, or recorded incorrectly, either due to human error and miscommunication (Nouraei et al. 2016; O'Malley et al. 2005), or because of systematic recording practices (Fawcett et al. 2019; Martin et al. 2024). On a fundamental level, there is no such thing as objective data, because which data is even collected and how it is measured is shaped by human values and decisions (Pine & Liboiron 2015).

These data quality issues will in turn affect models trained on such data. Sambasivan et al. (2021) define a "data cascade" as a series of data issues that compound over time to cause significant negative effects downstream. These original issues can be with data collection, initial analysis and cleaning, or they can be more fundamental, such as a misalignment of the purpose of the original data with the purpose of the analysis. At best, AI models from poor data lead to abandoned projects or expensive and time-consuming repeats of the data collection; at worst, operationalising such models could actively cause harm. Blindly applying AI to data without considering that data's provenance is unhelpful.

There are also practical issues associated with implementing AI technologies in practice. Cutting-edge deep learning models require substantial quantities of data to be effective, which as mentioned, is difficulty to gather in one place. Many healthcare systems lack the digital infrastructure to implement AI technologies; in the NHS, this includes the often poor computer infrastructure in NHS trusts, the considerable inconsistencies that exist between trusts, and the increased workload on already overstretched IT teams (Fazakarley et al. 2023). Proper and meaningful implementation requires the in-house capability to procure and implement systems; the resources to support their ongoing maintenance; and the know-how to select appropriate solutions, especially given the currently haphazard state of regulation and support (Karpathakis et al. 2024). More evidence is needed on the cost effectiveness, time savings, and resource usage of implementing such technologies (NICE 2023a,b)

Spending money on new data and AI technologies is therefore not a silver bullet that will instantly reform the healthcare system. Aggarwal et al. (2022; 2024) caution against an over-reliance on technological innovations, arguing that many of the key pressures currently facing the NHS are operation and societal: the large backlog of cancer patients awaiting diagnosis,

the challenges in workforce capacity, and future changes in demographic structure and widening society inequalities. An emphasis on cutting-edge technologies risks ignoring the social and economic contexts that influence healthcare outcomes, and the healthcare systems that deliver these interventions. Meaningful use of data needs to practically consider the system it operates in, and how it is contributing to improved patient outcomes.

2.2 *Artificial intelligence in healthcare*

The idea that AI could be applied to healthcare scenarios is almost as old as the field itself. The origins of AI as a discipline lie in the 1950s and '60s, when many of the basic ideas that still underpin both symbolic and neural methods were established. However, the relatively limited computational power of the time meant that many bold claims were never achieved (Kautz 2022; S. Russell & Norvig 2021).

In the 1970s and '80s, focus shifted to knowledge representation: the idea that an expert such as a doctor or engineer's knowledge in a particular domain could be encoded into an "expert system" was extremely popular (Haigh 2024b; Kautz 2022; S. Russell & Norvig 2021). These methods involved encoding knowledge into sets of facts and rules that could be reasoned over to make decisions (Shortliffe 1986). These approaches saw mixed success: they captured rules well, but probability and uncertainty poorly, and the sheer scale of work required to transform entire domains into sets of rules without error was in many cases untenable. Many were successful on a technical level, but did not achieve widespread adoption as a result of managerial and technical factors (Gill 1995). In many cases, businesses began to use these systems, but struggled to maintain the skills and staff required to keep systems up-to-date (Haigh 2024b; Kaul et al. 2020).

In the medical field specifically, progress slowed due to the difficulties of constantly updating the knowledge, the challenge of learning multiple systems for different diseases, and the reluctance of doctors to trust automated diagnoses. Whilst few complete expert systems made it into everyday use (Heathfield 1999; Wyatt & Spiegelhalter 1991), many of their principles were adopted, and continue to be used, in narrower *decision support systems* which focus on particular diseases or scenarios and combine logical rules with data from patients' record to aid in decision-making (Gordon 1996; Sutton et al. 2020).

The modern era of AI is often considered to have begun with the release of the AlexNet convolutional neural network in 2012, which dramatically outperformed traditional computer vision algorithms and popularised many-

layered neural networks. Since this point, *deep learning* (LeCun et al. 2015) has been by far the dominant paradigm in AI, and has received mainstream attention with the release of a variety of publicly available models, most notably OpenAI's ChatGPT large language model (LLM). At the time of writing, deep learning and LLMs are considered to represent the state of the art in AI.

Histories of AI typically describe the history of the field as a pattern of repeated "boom and bust" cycles, alternating between periods of intense progress and optimism (the three "AI summers" described above) and periods of disappointment and disillusionment (Cordeschi 2007; Kautz 2022; S. Russell & Norvig 2021). This characterisation of AI history as being divided into "summers" and "winters" has also been challenged. Whilst public interest, funding, and commercialisation of AI (in short, "hype") has wavered (Agar 2020; Haigh 2024a), many significant advances were also made during the so-called "winters" (Haigh 2023).

The present level of interest and investment in AI is undoubtedly unprecedented (Floridi 2024). However, whilst AI has achieved undoubtedly impressive results on a wide range of medical tasks in the academic literature, this does not guarantee successful translation to care scenarios, or meaningful improvements for the average patient's experience. In 2016, a noted AI pioneer infamously commented that "people should stop training radiologists now"; eight years later, radiologists — along with other medical professionals — still find themselves very much in demand. It is clear that there exists a significant gap between the state of research in medical AI, and real clinical practice in working healthcare systems.

This can be attributed to a number of factors. Firstly, there is uncertainty as to how well AI actually performs in the real world. A surprising number of such studies do not report results according to guidelines, and often make bold recommendations for clinical use without externally validating them in such scenarios (Andaur Navarro et al. 2023; Dhiman et al. 2023). It has been suggested that a large number of prediction models report inflated performance estimates (Kapoor & Narayanan 2023; N. White et al. 2023). Roberts et al. (2021), for example, examined 415 papers purporting to detect COVID-19 from imaging data, and found that none of them were suitable for clinical use due to methodological flaws and underlying bias. As Karpathakis et al. (2024) put it, "evidence of statistical accuracy is not evidence of clinical or operational effectiveness".

Modern AI also requires colossal quantities of data to train and test, which, as already discussed, are often not available in the NHS's fragmented data environment. Progress has been made on this front, with several projects

aiming to pool this data at the regional or national level (DHSC 2022; J. Morley & Zhang 2023; Nab et al. 2024), but the problem returns at the international scale, where different national datasets use inconsistent formats and are effectively incompatible. Healthcare data is not infinite: it is human-generated, substantially smaller than the volume of data available in other domains, and not guaranteed to grow at the same rate (Villalobos et al. 2024). The AI-aided generation of synthetic data is sometimes proposed as a solution, but research indicates that this is of limited use at scale: the accumulation of small errors means that at each stage, models are trained on worse and worse data, eventually leading to a complete collapse in quality (Shumailov et al. 2024).

This increasing scale applies not just to data, but also to computational power and expense, which is increasingly making model training the preserve of major corporations, and inaccessible to the academic community (Owens 2024). This requirement to store enormous volumes of data, train models for weeks on end, and manufacture and dispose of specialist hardware also creates a substantial environmental impact (Jay et al. 2024; Lucivero 2024; Strubell et al. 2019). The alternative argument, of course, is that there may well be sustainability benefits to be gained if AI can help to improve efficiency of medical treatment (Doo et al. 2024), but any such impacts are difficult to quantify, and any study that attempts to put a number to emissions relies on guesswork. Regardless, a truly useful data-driven or AI tool needs to be workable within the context of an increasingly resource-constrained healthcare system.

Finally, the increasing focus on large-scale "everything models", trained on enormous amounts of publicly available data, does not necessarily serve medical applications well. General-purpose models are usually outperformed in medical contexts by smaller, task-specific models (Brown et al. 2025; S. Chen et al. 2024; Lehman et al. 2023). The question-answering tasks that language-based models are designed to excel at are not representative of real-world use cases, which creates a gap between impressive "state of the art" performance on benchmarks, and the performance on metrics that really matter to healthcare systems (Wornow et al. 2023).

AI is commonly divided into two distinct halves. *Symbolic* AI incorporates methods based on human-readable representations of processes, such as logical rules and ontologies, whilst *sub-symbolic* or *connectionist* AI is based on correlations between input and output variables, and encompasses statistical learning methods such as deep learning and language models (Goel 2021; Ilkou & Koutraki 2020). Both paradigms have their advantages and disadvantages. Symbolic AI, being based on logical rules, can reason about

problems in predictable ways, and use explainable steps to do so. Relying on facts, rather than observed patterns, means that they do not require large amounts of training data. However, these rules require human effort to curate and maintain, limiting scalability.

Sub-symbolic methods, by contrast, rely on observing the statistical properties of a dataset and establishing correlations between variables, which makes them far more resilient to noise, but also entirely dependent on their training data, and far more opaque, especially in the case of deep learning methods which can make use of millions of individual weights. This reliance on statistical patterns means that they struggle to perform logical reasoning (Gerber & Eybers 2025; Shojaee et al. 2025). Language models, for example, create the *appearance* of logical reasoning by mimicking human writing, but without any particular regard for truth, instead assuming that correlation and frequency in training data corresponds with factual accuracy (Hicks et al. 2024). Within healthcare, this reliance on the training data is a fundamental issue, severely limiting models' generalisability, and introducing bias according to the particular patient population used.

Ideally, useful AI in a healthcare scenario should combine the advantages of both: sub-symbolic methods' ability to quickly and flexibly identify patterns in enormous datasets, with symbolic methods' ability to support decision making with known facts and logic, linking findings to strictly define concepts and following logical mechanisms of action.

2.3 *Clinical pathways*

In healthcare, a pathway is a particular "route" through identification, diagnosis, treatment and follow-up, usually specified in a guideline or policy that aims to standardise care for a particular condition or group of patients. A patients' routes through the healthcare system matter: the referral route taken by a patient, and the time taken to reach diagnosis or treatment, can have a measurable effect on outcomes (Neal et al. 2007, 2015). Pathways therefore exist to standardise these routes and improve outcomes. In the UK, the best known clinical guidelines are those published by the National Institute for Health and Care Excellence (NICE): their guidelines on colorectal cancer (NICE 2015, 2020) provide an idea of what the patient pathway looks like for a patient with colorectal cancer. Patients presenting with serious symptoms are supposed to receive an urgent referral to a specialist, and be seen within two weeks — this is known as the two-week wait (2WW) pathway. Patients who present with milder symptoms should first be offered a faecal immunochemical test (FIT), which tests for the presence of blood in the stool,

and referred through the 2WW pathway only if the test meets a certain threshold. Following a positive diagnosis, the first step for colon tumours is generally surgery, with laparoscopic ("keyhole") surgery the preferred option. This is typically followed by *adjuvant* chemotherapy, chemotherapy applied after the initial treatment to minimise the chances of recurrence, with three particular drug regimens recommended. In more severe cases, chemotherapy might occur before surgery, and in cases of bowel obstruction, stenting is an option before surgery, but only where the intention is for palliative rather than curative care. The pathway for rectal tumours is similar, with additional options for pre-operative radiotherapy in addition to chemotherapy, but no mention of stenting. In cases of metastatic cancer, tests for certain genetic mutations are recommended, before several different options for chemo- or immunotherapy, eventually followed by palliative care if necessary. This example demonstrates that the pathways recommended by guidelines — even the broadest guidelines decided at a national level — are rarely straightforward paths, but are complex sets of branching and intersecting routes that depend on many variables.

Pathways are not static, and will evolve in the light of new evidence and changing patient needs. For example, FIT is commonly recommended in colorectal cancer pathways as a "rule-in" factor where a patient does not otherwise qualify for a referral, but during the COVID-19 pandemic many providers attempted to ease pressure on hospital capacity by extending its usage to *all* symptomatic patients (Loveday et al. 2021). Occasionally, entirely new pathways are introduced from scratch. Cancer patients can sometimes experience symptoms that are clearly concerning, but not specific to any particular cancer site, which has can lead to delays and require tests to be repeated as they are re-referred along different pathways; in recent years, new pathways have been specifically devised for patients with these "low-risk but not no-risk" symptoms, in order to address these inefficiencies (Nicholson et al. 2018).

Ultimately, however, the guidelines that describe and promote pathways are just guidelines. Their interpretation and implementation will vary according to local needs, and clinicians will also incorporate their judgement, expertise, and opinion in making decisions.

2.3.1 *Evaluating clinical pathways*

Because different pathways have been associated with different outcomes, a key research objective is evaluating their effectiveness in terms of both clinical and economic outcomes. The implementation of pathways has been

associated with faster diagnosis, reduced complications, and reduced costs, amongst others (Neal et al. 2014; Rotter et al. 2010). However, not all studies have found consistent results: effectiveness varies in different disease areas and in different patient groups (Allen et al. 2009). A set of guidelines is only as good as the evidence it is based upon — it is therefore important to conduct evaluations of clinical pathways in practice, to establish whether a patient's adherence to a particular pathway is actually associated with any clinical benefits in terms of their outcomes.

Lee et al. (2019) systematically review the approaches used to evaluate clinical pathways. The majority of evaluations use a pre-post study design, meaning that they compared cohorts of patients who were treated before and after the implementation of a particular pathway. Most frequently, this meant that the pre-intervention cohort was studied retrospectively from historic patient data, and the post-intervention cohort was studied prospectively by observing patients following an explicitly designed pathway. The second most common design was the case-control study, where control and case groups from the same period in time were compared, either prospectively or retrospectively. Most studies used classical statistical tests such as Student's t-test, the χ^2 test, or Fisher's exact test to compare the two groups. Nine studies estimated covariate effects using a regression model; only one applied Kaplan-Meier survival analysis. 90% of articles identified some improvement in patient outcomes, including shorter length of stay, reduced variation in clinical processes, and fewer adverse events, although the authors acknowledge that publication bias likely led to an overestimation of effectiveness.

This review provides a useful window into the methods currently used to evaluate pathways, but it is very limited in that it only considers studies that conducted economic evaluations of pathways, which eliminated 473 out of the 528 initially identified. Economic metrics are only one way to evaluate patient pathways, with patients' health outcomes being another — perhaps greater — priority.

Rotter et al.'s systematic review (2012) is older but more thorough, and considers four different study designs: randomised controlled trials, controlled clinical trials, controlled before and after studies, and interrupted time series analyses. They observe that more than 70% of studies were pre-post comparisons of yearly cohorts — confirming Lee et al.'s findings — but argue that such studies should be excluded entirely from consideration as they are likely to be misleading, since the time difference is likely to introduce a number of confounding factors such as changes in case mix, changes in hospital policy, or more general hospital quality improvement. This sup-

ports the conclusion that on the whole, there are relatively few high-quality evaluations that investigate the effectiveness of clinical pathways.

A handful of attempts have been made to analyse pathways in alternative ways, for example by assigning patients scores based on their similarity to a particular pathway. Forster et al. (2020) develop a measure of an individual patient's concordance with a pathway. They identify the steps most essential to survival, in consultation with domain experts, and assign them specific time intervals and a maximum permissible number of encounters. From these five key events, they establish four possible sequence groups that they could occur in, three of which are considered concordant with the pathway. Based on this definition, the cohort is then divided into concordant and non-concordant patients, and logistic regression used to identify predictors of concordance. The measure of concordance is relatively crude, including only 5 elements of care, with no weighting according to importance or association with outcome measures. This raises a key challenge of pathway research: pathways are often described in ambiguous or unclear ways. This study also highlights the high level of clinician engagement required, which makes generalisation of this technique to other diseases time-consuming and expensive. Additionally, only stage II and III colon cancer are included in the study, as "the pathway maps for those two groups were felt to be the most clearly specified".

Van Zelm et al. (2018) also describe a protocol for evaluating pathways, which involves dividing a pathway into five components — intervention, context, implementation, mechanism of impact, and outcomes — each evaluated with either quantitative and qualitative methods, or both. Adherence is quantified according to the percentage of these interventions the patient receives.

The majority of analyses of clinical pathways therefore rely on dividing patients into case and control groups of patients who followed the pathway, and patients who did not. Whilst this analysis might be appropriate for a situation such as a randomised controlled trial, where patients' treatment pathways can be prospectively assigned and rigorously enforced to ensure in-group homogeneity, an increasing proportion of studies now consist of retrospective analyses of routinely collected data. It is therefore likely that binary "pathway" and "non-pathway" groups conceal significant heterogeneity, and such analysis would be better conducted by evaluating patients in more granular groups, according to multiple dimensions of possible divergence from the pathway.

2.3.2 *Summary*

In summary, patient pathways are a widely used tool to standardise and improve healthcare outcomes. It is clear that no "gold standard" method for comparing patients to a particular pathway exists. Historically, evaluations have compared the outcomes of patients before and after a change in policy, but this is vulnerable to any number of confounding factors. Any evaluation of clinical pathways also needs to take into account that every patient is an individual, that deviations from a recommended pathway often happen for a reason, and therefore that enforcing or encouraging complete compliance with a pathway in all cases is neither feasible nor useful.

2.4 *Analysing pathways*

There clearly exist a substantial number of research questions around the delivery and effectiveness of clinical pathways, but a relatively small set of methods are used to answer these questions in the established healthcare literature. A number of alternative methods have been proposed as solutions, most notably process mining and electronic phenotyping, but there is also relevant research on similar problems in other medical scenarios. This section examines a variety of existing methods that have previously been used to analyse pathways, histories, journeys, sequences, or trajectories in some way; whether by characterising them, comparing them, or identifying similar instances.

2.4.1 *Electronic phenotyping*

EHR systems are a rich source of data, and they have proven a powerful resource in recent years for planning and research. However, all medical records fundamentally rely on human data entry. This labour-intensive coding process, combined with the sheer number and complexity of stages in a patient's treatment introduces numerous sources of accidental error, making inaccuracies not just possible but inevitable. Miscommunication between patients, clinicians or coders, lost or misplaced paperwork, the level of training and experience of using EHR systems, and transcription errors are all potential issues (O'Malley et al. 2005). Additionally, deliberate practices such as "up-coding" and "coding inflation", i.e. choosing the most expensive code or adding extra secondary codes to maximise reimbursement, have been reported in the UK (Fawcett et al. 2019).

A UK-based study of over 8,000 discharge reports found that when audited, 55% of records required at least one change, and 16.8% required a

change to the primary diagnosis (Nouraei et al. 2016). Another study looking specifically at endocarditis found that estimating incidence from ICD-10 codes lead to a twofold overestimate in cases. However, critical selection of code inclusion criteria by domain experts — as opposed to simply including all codes containing the term "endocarditis" — did improve accuracy, suggesting that some process of human code validation is necessary for EHR-based studies to be meaningful (Fawcett et al. 2019). This in turn creates additional problems, where differences in opinion lead to different studies using different definitions of the same disease, although repositories of standard codelists have been developed as one solution to aid reproducibility (Denaxas et al. 2019).

Because of these inaccuracies in clinical coding, identifying particular cohorts of patients generally involves the analysis of multiple data points in addition to just diagnosis codes. This process of identifying patients with particular characteristics of interest is known as *electronic phenotyping* (Banda et al. 2018). Phenotypes can vary widely in complexity and specificity, from "patients with x disease" to "patients admitted in time window $t_1 ... t_n$ with x, y and z diseases with symptoms a and b and outcome o". The ability to retrieve cohorts of patients according to these criteria is vital for almost any study type that relies on EHR data, including epidemiological studies, genome-wide association studies, predictive modelling, and clinical trial recruitment. Banda et al. define three main approaches to electronic phenotyping. The first, the traditional rule-based approach, involves specifying inclusion or exclusion criteria based on codes or values, as in the aforementioned examples, and has been used since at least the 1990s. Over time, the addition of further data sources such as clinical text keywords and medication data has further improved precision, although retrieval performance can vary widely depending on the disease. In particular, more complex queries involving multiple diseases and fragmented datasets can exhibit poor sensitivity. These increasingly complex definitions can come at a price: for example, K. I. Morley et al. (2014) define an atrial fibrillation phenotype that incorporates 286 codes across four coding systems, but the process required an extremely time-consuming process of iterative expert review.

The second group includes the increasing number of approaches that make use of additional information available from unstructured sources such as clinical notes, letters, and reports, originally with pattern matching and in recent years with natural language processing (NLP) techniques. Phenotypes that combine traditional rules with NLP techniques have demonstrated strong performance, and are now widely used. The third and final group of phenotype methods are the emerging machine learning-supported

approaches, most significantly the introduction of "high-throughput phenotyping", which marks a shift away from hand-crafted disease-specific phenotypes, and towards scalable and unsupervised generation of many phenotypes using machine learning (ML) approaches. Validating phenotypes remains challenging however, with many approaches relying on clinician review (Kho et al. 2012).

The idea of identifying patients who follow particular pathways can be viewed as a form of phenotyping, but one that is defined in terms of procedures and interventions rather than the more common diagnoses and symptoms. The parallels between electronic phenotyping and clinical pathway analysis have previously been observed (Dagliati et al. 2017), since they share the key objectives of identifying distinct cohorts and representative features for specific diseases. Pathway analysis tasks, in particular clustering, can be viewed as effectively a form of electronic phenotyping that places a greater emphasis on procedures and interventions rather than symptoms and diagnoses, and generally being longitudinal rather than cross-sectional in nature. Therefore, it is likely that some of the methods used in electronic phenotyping might, with some modification, be useful components in a pathway analysis toolkit, although since the data comes from the same sources all of the drawbacks of EHR data will still apply.

2.4.2 *String metrics*

The most straightforward approaches to comparing pathways represent them as simple strings, with each letter representing a particular event. Williams et al. (2014) demonstrate that basic string metrics, such as the edit distance or longest common subsequence (LCS), can be used to match patient trace strings to pathway strings with very high accuracy. However, the method is demonstrated on a simple pathway with only two decision points, and there is no discussion of the granularity of the original dataset or the data preparation process, meaning that it is not clear which information was present in the original dataset and how decisions around exclusion or inclusion of original events were made. Vogt et al. (2018) similarly demonstrate that a distance measure based on the LCS metric can be used to cluster patient pathways and identify distinct subpopulations, as well as identify the most effective sequences for optimising outcomes. In both cases, it is unclear how well such an approach would scale to, a modern cancer pathway with significantly greater complexity and potentially thousands of possible different events.

Both of these examples treat procedure codes as unstructured, discrete concepts without acknowledging that some procedures are inherently more

similar than others; Aspland et al. (2021) address this by using a modified version of the Needleman-Wunsch algorithm, a long-established algorithm widely used in the field of bioinformatics to align genetic sequences. In their modified version, three additional rules are introduced to the algorithm: some activities are not allowed to substitute for others, where an activity is intrinsically "too different" to the one recommended; secondly, some activities are designated as members of the same group, making them interchangeable without penalty; and finally, activities can be weighted according to their importance in the care process, with the weights subjectively defined according to clinician input. In terms of accuracy, this approach matched classic text metrics similar to those used by Williams, but the incorporation of domain-relevant constraints on the matching process arguably make the resulting clusters more clinically meaningful, and a better reflection of the true nature of clinical pathways.

2.4.3 *State sequence analysis*

State sequence analysis (SSA) is a methodology that originates in the social sciences, originally developed for tracking career or educational patterns. Le Meur et al. (2015) are possibly the first authors to apply this method in a healthcare context, to compare prenatal care trajectories. Roux et al. (2019) provide a deeper dive into the methodology, and apply it to multiple sclerosis. In their approach, a care pathway is a sequence of states representing levels of care consumption — for example, a period of high care contact representing more intensive treatment, followed by a period of lower care consumption representing stable disease. The unit of time can be a standard unit such as days or months, but it can also be specific to a particular scenario, for example the trimesters of a pregnancy. An "alphabet" of possible values for each state needs to be created — Roux et al. use five possible levels of healthcare utilisation; these are found by summing the total annual number of healthcare consultations and admissions, and splitting them into four quartiles plus a zero group. Following this, the similarity between sequences is measured, usually based on string metrics such as the Levenshtein or Hamming distance, LCS, or simply the number of matching subsequences. The resulting dissimilarity matrix can be used as input for a variety of ML tasks, typically clustering.

The SSA approach is relatively flexible in its application, since the states do not necessarily have to represent care consumption — their meaning can be chosen by the researcher. For example, a later study by Le Meur et al. (2019) uses states representing one of five possible treatment modalities,

plus waiting and death. Savaré et al. (2023) count the number of discrete treatment types given in a week. Biggin et al. (2023), investigating outpatient appointments, consider two possible state encoding schemes: a binary system, where time is quantised into months with a value of either *appointment* or *no appointment*, or alternatively according to an alphabet of five possible values representing the outcomes of an appointment, ordered simply by appointment number rather than proportional to time.

Counting the number of consultations in a given time period — as used by Roux et al. (2019) and Roth et al. (2022) — is an effective, but relatively simplistic view of a healthcare system. A future extension might consider weighting each consultation based on, for example, its intensity, its perceived importance to the treatment process, or to its resource usage from the provider's perspective. Ultimately, SSA is a useful approach for examining healthcare service usage from an economic or social perspective, but it reveals a limited amount about the actual treatments offered from a clinical perspective. There is therefore room to adapt these methods to consider the specific sequences of treatments offered, and examine the relationship between these sequences and clinical outcomes.

Comparative studies have indicated that SSA provides results comparable to latent class analysis (LCA) (Barban & Billari 2012; Han et al. 2017; Mikolai & Lyons-Amos 2017), but does not require specification of a model. The specific distance metric chosen can affect the output: Studer & Ritschard (2016) compare several different distance measures applied to life trajectories, highlighting that there are few "incorrect" distance measures and that the the choice depends on the focus of the study. For example, researchers might desire easy identification of changes in sequencing, sensitivity to small perturbations, or changes in timing, all of which might suit different measures.

Representation aside, SSA approaches are built on string metrics and therefore share their disadvantages. Most measures based on sequence similarity fail to take into account semantic difference: where there are many possible states in a sequence, it is likely that some states are more similar than others, and approaches that rely on counting the number of edits will not take this into account. Rivault et al. (2017) propose a solution that modifies the LCS measure to take into account concept similarity, as quantified by the number of nodes separating the two concepts in an ontology or hierarchy.

2.4.4 *Process mining*

Process mining (PM) is a family of methods focused on extracting insights from event logs — chronological records of historic activities. PM is classically divided into three areas: *discovery*, in which event logs are analysed to discover process models; *conformance checking*, in which the behaviour of event logs are compared to existing process models; and *enhancement*, in which existing process models are extended and improved (van der Aalst 2012). Despite the success of PM in business and industry, it is often observed that their results can be more mixed on medical data, with algorithms being poorly equipped to deal with unstructured processes, and strongly influenced by noise, incompleteness and the sheer number of process variants (Kaymak et al. 2012; Lang et al. 2008; W. Yang & Su 2014).

Healthcare processes evidently present a number of unique challenges. They are characterised by four key properties: they are highly dynamic, changing as new interventions and technologies are introduced; they are highly complex, relying on large amounts of data and unpredictable events; they are increasingly multidisciplinary, requiring coordination between multiple specialist units in an organisation; and they are ad hoc, often modified and interpreted according to individual preference and professional judgement, making processes highly variable, non-repetitive and non-deterministic (Rebuge & Ferreira 2012). Healthcare processes therefore create data characterised by incomplete and noisy signals, high levels of variation in processes, and an abundance of exceptional behaviours, many of which should ideally be captured rather than disregarded. Together, these key obstacles mean that the most common PM approaches do not always produce useful results when applied to healthcare data and problems.

This does not mean that PM does not work on healthcare data at all. There has been some success applying PM to disease trajectories, i.e. diagnosis-only data (Kusuma et al. 2020, 2021), but a study of patient pathways naturally needs to include a much more diverse set of data, covering diagnostic and treatment processes. Mans et al. (2008) analyse stroke care across two different hospitals and create two separate process models, allowing the two centres' practices to be compared. This produces useful and interpretable insights, for example, the authors found that one hospital clearly performs hypertension therapy earlier and more frequently.

Rebuge & Ferreira (2012) address these drawbacks by proposing a new PM methodology that emphasises infrequent behaviours and process variants. Sequence clustering is performed by generating Markov chains with random transition matrices, and iteratively assigning traces to the most likely

source chain to have produced it, until clear clusters of high support (the most common processes) and low support (detected variants and outliers) emerge. However, this technique is only tested on a relatively small radiology workflow, a process of six steps undertaken as part of one episode of care. Whether or not this method would generalise to an entire treatment pathway, which would likely be far more complex and contain greater variation, is untested.

In summary, it is clear that simply applying off-the-shelf PM techniques to clinical datasets does not guarantee useful results, often leading to either invalid or partial models, or models which do not reflect reality. Whilst PM is domain-agnostic in principle, numerous studies have indicated that the distinctive nature of healthcare processes creates real practical challenges. Some healthcare PM applications have generated useful results in small case studies, but the methods are yet to fully account for the dynamic and complex nature of pathways and prove themselves on large-scale multidisciplinary pathways.

2.4.5 *Machine learning*

A number of studies have sought to use machine learning (ML) techniques to analyse pathways. Paik et al. (2019) extract "diagnosis trajectories", i.e. commonly co-occurring pairs of diseases, and use an algorithm to concatenate them into multiple graphs of the most common disease sequences. Much like some of the previous disease trajectory approaches in PM, this is useful but does not incorporate the range of event types needed to gain a clear understanding of clinical pathways. Pokharel et al. (2020) propose structuring patients as a "temporal tree", a set of trees each describing the events at a particular time point. These trees are then transformed into a string representation via either breadth- or depth-first search, and this string is converted to a low dimensional vector space through word embedding algorithms. Given that the actual analysis is performed on a string representation of the tree, rather than the tree structure itself, this could be viewed as an extension of the simple text metrics described in Section 2.4.2, but using more advanced NLP techniques in place of rule-based algorithms.

Some approaches use more complex ML techniques such as neural networks. Castela Forte et al. (2021) compare a number of different ML algorithms, and find that deep embeddings learned through neural networks generally outperform traditional methods for clustering. Carr et al. (2021) describe an autoencoder approach to patient clustering. Instead of ignoring information that does not predict future events, as is common in other cluster-

ing approaches, the patients' full trajectory is retained to prioritise clinically interpretable clusters. Rather than using standard clustering methods on the output vectors, the process is incorporated into the neural network by adding a clustering layer, allowing cluster assignments to back-propagate and optimise the learned embeddings for clustering purposes. The approach combines predictive (outcome-centric) and unsupervised (pathway-centric) approaches, allowing the weighting to be changed to prioritise one or the other, which is an important consideration when analysing pathways. A system for analysing pathways should ideally have some such flexibility, but for most studies that involve associating pathways to outcomes and measuring different outcomes between pathways, including outcomes in the learning process is less desirable.

Van Smeden et al. (2018) urge caution when using clustering to find subgroups in data, in particular because the lack of ground-truth data in an unsupervised approach makes ascribing any meaning to these clusters difficult, leading investigators to rely on subjective interpretations and potentially fit their theory around the data. Clusters often represent dependencies between variables, rather than any true subgroups; it is therefore preferable to draw on existing theory *before* identifying subgroups to prevent the making up of new theory to fit the data. Alternatively, clusters should be evaluated in terms of their utility for prediction, to ensure that they truly correlate with the desired outcome. They also observe that using the same set of variables in a generalised linear model will almost certainly provide a more "personalised" prediction than treating patients within a cluster as homogenous.

Whilst outcome prediction approaches should therefore predict on a per-patient basis, clusters still represent an effective tool to aid human understanding and interpretation of data, and certainly have a place in identifying natural groups of patients who experience the same pathway.

ML approaches have therefore demonstrated significant potential for analysing sequences of patient interactions, and can be seen to outperform more traditional methods. However, a key issue with high-performance methods such as neural networks is interpretability and explainability: whilst this might seem a lower priority for applications that do not produce a prediction or contribute directly to decision making, pathway analysis will still benefit from a clear rationale of decisions made in order to generate useful insights into the reasons why certain pathways might share certain outcomes. The success of simple text metrics demonstrates how domain-relevant rules can produce insights that make sense and follow real-world

rules, whilst ML approaches show very high performance — the natural next step is an approach that combines the best of both.

2.4.6 *Alternative approaches*

Outside of these main approaches, there are a number of other methods which have been used in individual studies. Bettencourt-Silva et al. (2015) suggest a relatively simple "completeness score" for a patient, counting how many steps in the pathway have been completed for each patient. Huang et al. (2014) use latent Dirichlet allocation (LDA), a type of probabilistic graphical model, which assumes that all possible treatment behaviours can be represented by a much smaller number of simple behavioural "topics", each with its own probability distribution. By combining these topics with the original patient traces, similarity can be measured. A particularly novel idea is presented by Nguyen-Duc et al. (2021), who encode patient pathways into image form, and use deep learning for prediction, emphasising the potential interpretability of such a technique.

Perer et al. (2015) analyse patient traces using sequential pattern mining approaches, specifically the Sequential Pattern Mining with bitmap representation (SPAM) approach. By using a bitmap-style representation and representing event sequences as binary codes, SPAM allows patterns to be detected using simple binary and/or operations, making it highly efficient. However, a key drawback highlighted by the authors is that SPAM does not have the capability to incorporate certain constraints, making it hard to look at a particular domain-relevant time window.

There is also a significant body of research on disease trajectories that aims to identify frequent and recurring patterns of disease progression from datasets; in principle, these methods intended to identify common sequences of diseases could well be adapted to identify common sequences of procedures. Thygesen et al. (2022), for example, identify COVID-19 related phenotypes representing different stages of disease based on events including testing, hospital admission, and intensive care admission. As well as measuring the proportions of patients exhibiting each phenotype, they also count the numbers of patients transitioning between each stage — effectively creating a directly-follows graph of the style popular in process mining research. Hu et al. (2019) use a similar approach to identify precancer disease routes, whilst Lademann et al. (2019) incorporate symptom data for more detailed stratification.

Typically, these models are constructed by identifying pairs of commonly co-occuring diseases that typically appear in a particular order. Some studies

have combined individual pairs of trajectories in larger trajectory networks or graphs (Siggaard et al. 2020). However, a key issue with this typical approach is that it does not confirm whether these longer trajectories actually occur in the data. Kusuma et al. (2021) present process mining as a potential solution to this, but find in a literature review that only handful of papers have used PM for the specific problem of disease trajectories.

2.5 *Knowledge graphs*

It is increasingly suggested that the future of AI will be in *neuro-symbolic learning*: the combination of advanced neural techniques, focusing on learning vector representations, with older logical or symbolic techniques, focusing on representing facts and expert knowledge. A major concept amongst these symbolic methods is the *knowledge graph*, a family of methods that combine graph-like data models with semantic relations and reasoning. Knowledge graphs (KGs) are closely related to, and often work alongside, *ontologies*, formal representations of the concepts and relationships between them in a particular domain. However, an exact definition of the term can be hard to come by. The term was popularised by Google in 2012 to refer to semantic enhancements in the Google search engine (Singhal 2012), although the underlying ideas have a long history: the concept of graphically representing knowledge was popularised in the 1950s, logic programming encouraged the closer integration of data with knowledge in the 1970s, and the 2000s saw the development of the semantic web and linked data fields (Gutierrez & Sequeda 2021).

Ehrlinger & Wöß (2016) argue that two major factors have led to confused definitions of KGs: firstly, that Google's original announcement is widely cited despite the fact that it lacks any explanation or technical details, and secondly that "knowledge graph" is commonly used as a synonym for *knowledge base* or *ontology*. They therefore define a KG as something that *acquires and integrates information into an ontology and applies a reasoner to derive new knowledge*. Here, a KG is distinguished from an ontology by the fact that it employs reasoning to a set of ontological facts to generate new knowledge.

In practical terms, a knowledge graph typically takes the form of a database structured in a graph form, where nodes represent real-world entities, and the edges represent the semantic relationships between them. The core unit of information in a KG is the *semantic triple*, a set of three entities that encode the relationship between two entities in the format *(subject, predicate, object)*; for example *(patient, has diagnosis, colorectal cancer)*. These

triples can be represented in a number of different graph models, including as a directed edge-labelled graph, in which relationships carry both a direction and a label; a heterogeneous graph, containing nodes of different *types*; or a property graph, where edges, as well as nodes, can have associated properties (Hogan et al. 2021).

KGs are extensively discussed in this thesis for two reasons. Firstly, KGs are very closely intertwined with ontologies, which are themselves widely used as standards for the structure and reporting of healthcare. Secondly, a large body of work exists on using KG representations of knowledge and data to create vector representations of knowledge, which can be used for a variety of different downstream applications. KG methods are therefore a natural way to interact with the knowledge stored in EHR data and support its reuse, whilst also preserving its meaning and semantics.

Many authors have argued that graph database systems are well-suited to managing EHR data. From a purely computational perspective, an EHR graph database is faster and more intuitive to query, and they facilitate the integration of different sources of data, which is a key requirement in EHR contexts (Dwyer et al. 2023; Stothers & Nguyen 2020; Yoon et al. 2017). In particular, a graph-based approach allows for patient data to be closely integrated with widely used ontologies such as SNOMED CT, which follow similar graph structures (Campbell et al. 2015). However, whilst some authors have advocated moving EHR systems away from traditional relational systems and towards graph databases, the effort, expense, and time required makes a wholesale migration on any local or national scale exceptionally unlikely in the near future, and is certainly outside of the scope of this thesis.

KGs are one of many methods for knowledge representation, and whilst their flexibility and expressiveness makes them useful tools, they also come with disadvantages. Whilst they represent a ground truth source of information, this information needs to be curated: either by humans, requiring time and attention, or automatically, which requires oversight and quality assurance. They also need to be able to be updated as the facts they are based on change and evolve with scientific progress. It is also challenging to represent multi-modal data sources, for example images or numeric data, without extracting features from them in a lossy manner (J. Chen et al. 2023). Whilst KGs themselves are flexible, downstream uses such as low-dimensional embeddings derived from them do not always respect logical axioms or effectively model more complex relation patterns. Particular care needs to be taken to use the correct model for the correct application, as I will explore in more detail in Chapter 4. The remainder of this section

introduces key KG concepts in more detail, and discusses the current state of KG methods in healthcare research.

2.5.1 *Reasoning*

A key feature of knowledge graphs is that the logical statements that form them can be reasoned over to produce new knowledge not explicit in the original data. This new knowledge can be derived from a KG either *deductively*, where reasoning is applied to existing facts to derive new ones based on logical rules, or *inductively*, where patterns are generalised from the facts to generate new probable facts (Hogan et al. 2021). The most common application of knowledge graph reasoning is in link prediction, which is used to infer the relationships between entities, allowing incomplete data to be filled in or new relations predicted . This is commonly achieved through an embedding-based approach, in which KGs are transformed into a vector representation that preserves its structure, and predictions made based on entities' similarity within the embedding space (Bellomarini et al. 2020). Reasoning can be achieved either through either such an embedding-based approach, or by using sets of formal logical rules. Increasingly however, these two approaches are being merged to create new hybrid methods, for example by applying logical rules on top of learned embeddings as a post-processing step, or by integrating logical rules and ontological knowledge directly into the embedding learning process (d'Amato et al. 2023)

2.5.2 *Embedding*

A knowledge graph embedding (KGE) is a representation of a knowledge graph in a vector space, the aim being to preserve the graph's structure whilst simplifying tasks such as clustering, link prediction, or entity resolution. The process of converting a KG to a KGE generally involves three steps: representing a set of triples (h, r, t), applying a scoring function (to measure the plausibility of a proposed fact), and learning the representations of the entities and relations (an optimisation problem of maximising the plausibility score). This subsection discusses some of the different approaches to embedding, as well as how embeddings can be applied to generate new knowledge.

There exist a large number of KGE methods, and embedding methods are increasingly leveraging further information such as entity types, relation paths and logical rules to make embeddings that are more predictive. Most of these approaches work by representing entities as points in a vector space, and relations as some kind of operation within the vector space, which can

be described with a vector, matrix, tensor, or probability distribution. These representations are then evaluated in terms of plausibility using a scoring function, with facts observed within the KG scored higher than those not observed. Finally, an optimisation problem is defined to maximise the total plausibility of the observed facts (Wang et al. 2017).

KGEs broadly fall into two groups: translational distance models, and semantic matching models. Translational distance models exploit distance-based score functions, defining plausibility as the distance between two entities. Most famously this includes TransE, and its many extensions such as TransH and TransR. Translational distance models can however be simpler, for example the unstructured model (UM), a naive TransE which ignores different relation types; or more complex such as Gaussian embeddings, which consider entities and relations as random vectors drawn from multivariate Gaussian distributions, thereby allowing uncertainties to be taken into account. Semantic matching models, by contrast, use scoring functions that consider similarity rather than distance. RESCAL, for example, associates each entity with a vector to capture its latent semantics, with each relation being a matrix containing the pairwise interactions between latent factors. DistMult simplifies RESCAL by considering only diagonal matrices, but the trade-off is that it can only consider symmetric relations, which limits its usefulness in some scenarios. Neural network approaches are also possible. In its simplest form, a network such as a multi-layer perceptron, takes as input a vector representing either an entity or a relation. A given fact (h, r, t) is concatenated in the input layer, mapped to a non-linear hidden layer, and its score generated by a linear output layer.

When creating embeddings using any of these approaches, the initialisation of embeddings is usually random, from a uniform or a Gaussian distribution (Wang et al. 2017). Alternatively, some models will initialise their model with the result of a simpler model, for example as TransR starts with TransE embeddings (Lin et al. 2015).

Model training requires one of two assumptions to be made. Under the *open world assumption* (OWA) it is assumed that a KG contains only true facts, and an unobserved fact could be either false or missing. By contrast, the *closed world assumption* (CWA) assumes that any fact not present in the KG must automatically be false. This allows representations to be learned by minimising a function such as the squared loss, a more computationally efficient operation, however it of course rarely holds for large-scale real-world KGs, especially in a healthcare scenario (Wang et al. 2017). In a real-world scenario, the OWA is generally most realistic, but false triples are still required as training examples. The compromise is often to use a heuristic

approach such the *stochastic local closed world assumption* (sLCWA), in which false training triples are randomly generated, relying on the assumption that since the set of all possible triples not in the KG is so much larger than the set of triples within the KG, the chances of generating a false negative is so low as to be negligible (Ali et al. 2022).

Chang et al. (2020) identify knowledge representations as a weak area in the current literature. They argue that although a vast number of publicly available biomedical knowledge bases and ontologies exist, a lack of reliable methods for learning knowledge representations limits their usefulness in practice. Furthermore, much existing literature utilises network embeddings and graph embeddings as opposed to knowledge graph embeddings, meaning that semantically rich information is not being leveraged. KGEs therefore provide a promising path, but are relatively untested within the biomedical domain with no currently established best practice for training and comparing biomedical entity embeddings. KGEs have great potential in this area since healthcare concepts intrinsically contain rich latent information. Different procedures, prescriptions and conditions are intrinsically linked, for example some diseases will be more similar than others, which means that ML applications that use simple methods such as one-hot encoding to represent the presence or absence of a concept will miss out on this information (Choi et al. 2016).

2.5.3 *Applications in health and medicine*

KGs are an appealing way to represent biomedical knowledge for several reasons. Fundamentally, a graph-based structure is an intuitive method for complex networks of contextual information. These are widespread in health and medical contexts: this might apply to a sequence of clinical events, interactions between biological entities, or a patient's relationships with multiple comorbid conditions. In addition, analysing such problems generally involves integrating multimodal data from heterogeneous sources, which is significantly easier since graph databases typically rely on less rigid schemas compared to traditional database approaches. For this reason, there is a large body of work that uses KGs to solve problems in the biomedical and healthcare sciences.

Many approaches use straightforward graph structures to perform standard tasks such as outcome prediction, risk scoring or clustering. Khan et al. (2018), for example, construct networks of comorbid diagnoses for individual patients, then aggregate these into cohort-wide graphs, and compare these graphs across cohorts to identify sets of comorbidities that are more preva-

lent in diabetic than non-diabetic patients. Tissot & Pedebos (2021) create a graph of diagnoses, procedures, prescriptions, and demographic factors, and generate personalised risk scores based on a patient's neighbouring embeddings. The aim is to demonstrate that relatively simple embedding approaches such as TransE can produce excellent results when relevant, domain-specific constraints are integrated.

Pai & Bader (2018) describe the "patient similarity network" design, a graph where nodes represent patients and edges represent pairwise similarity for a given feature. They outline several advantages of this method, in particular the ease of handling heterogeneous data: any data type can be converted into a network as long as a similarity measure can be defined. This approach has been applied to both clustering and classification of patients, and the authors also argue for the interpretability benefits of being able to clearly visualise the decision boundary within the patient similarity space (Pai et al. 2019).

Xu et al. (2019) build a network of co-occurring disease pairs, and create a classification system that can predict the probability of a given disease occurring given a set of that patient's historic diseases. They also emphasise the interpretability benefits of a graph approach, since the system is able to generate "risk propagation" paths, which clearly illustrate which other nodes in the network have influence the final score. However, this work does not include any discussion of demographic variables, so it is unclear as to whether this approach generates any new non-trivial information that could not already be predicted based on, for example, age.

Outcome prediction in particular is a widely studied application of medical KGs, and many studies indicate that graph-based approaches can outperform more traditional methods. For example, Bean et al.'s KG of drugs, protein targets, indications, and reactions (2017) is able to predict adverse reactions to drugs, and outperforms logistic regression, decision trees, and support vector machines. Other studies demonstrate the advantages of using graph data and methods in combination with others. For example, a patient-level transcriptomics dataset enhanced with a KG of known protein-protein interactions produces new patient representations which enable improved classification performance compared to the original data alone (Bharadhwaj et al. 2021). Tong et al. (2022) combine temporal features in long short-term memory networks (LSTMs) with patient neighbourhoods in graph neural networks (GNNs) to predict hospital length of stay. The combined LSTM-GNN model outperforms an LSTM-only model, indicating that graph representations integrated into existing models can improve performance on certain tasks. Overall, previous studies have demonstrated that a KG approach to

prediction can generate new insights and improve performance, both on the data level - where ontologies can inject prior knowledge into observational data - and also on the level of methods and algorithms, where methods such as GNNs and embeddings can achieve improved performance.

Many approaches generate insights from existing data on biological entities, including proteomics, genomics, and metabolomics (collectively referred to as *multi-omics* or *omics* data). Vlietstra et al. (2020) describe constructing a graph of protein-protein and disease-protein predicates, and inferring trajectories between diseases where two diseases share a linked protein. Evaluation via expert review suggested true disease pairs can be detected with an AUC score of as much as 83%, depending on the reference set used. It is suggested that this approach might lead to the identification of new protein paths, leading to a deeper understanding of the mechanics of disease progression. Many approaches additionally combine this sort of data with knowledge extracted from published literature; for example, Vlietstra et al. (2017) create a graph combining multiple biomedical databases with publication abstracts, and extract biomedical concepts based on their connectivity to a small input set of concepts of interest, effectively allowing potential biomarkers to be detected based on their support in the literature. Similarly, Gogleva et al. (2022) combine genetic data with both information from existing KGs and literature mentions, and rank genes based on their relevance to a disease of interest. The authors frame this as a recommendation problem, demonstrating how existing KG techniques can be easily applied to domain-specific problems. Whilst these graph methods that analyse large volumes of multi-omics data generally cannot themselves provide direct experimental evidence for the relationships between genes and diseases, they are useful in sifting through huge volumes of data and narrowing the search space, in order to suggest priorities for bench research.

Despite these recent innovations in using graph methods for large volumes of health data, it has been observed that relatively little published work has investigated the potential of knowledge graphs on a smaller scale, for example representing and investigating an individual patients' data (Schrodt et al. 2020). Instead of constructing massive knowledge graphs linking entire diseases and concepts, an alternative approach might be to construct a KG centred around the single patient, thus integrating contextual information to better inform treatment. Such an approach would be useful in precision medicine applications, able to consolidate clinical and social context and recommend personalised treatment (Gyrard et al. 2018; Rastogi & Zaki 2020).

2.5.4 *Knowledge graphs and process mining*

Object-centric process mining is a more recent approach to PM that concerns itself with the description of processes from multiple points of view (van der Aalst 2023). Traditional PM assumes that a process model describes the lifecycle of a single object, and that each event refers to exactly one object of a given type. In reality, processes are rarely independent: the process of a candidate applying for a job, for example, is intertwined with those of other applicants, as well as a different hiring process from the point of view of the recruiter. Therefore, an activity may well involve different objects of different types. Berti et al. (2023) present a graph-based approach to feature extraction from object-centric event logs. The relationships between the objects in a log are described using an object-based graph, and objects are associated with numeric features, used as input for ML tasks.

It can similarly be argued that this also applies to patient pathways: a treatment pathway is undoubtedly a complex web of many processes that can be viewed from the point of view of the patient, the clinicians, pharmacists or from a purely logistical standpoint. It is also possible, however, that given the size of EHR datasets, an analysis that incorporates all the possible interwoven threads and processes could become combinatorially infeasible and practically uninterpretable. Each individual analysis therefore needs to judge the correct abstraction to make for a particular research question as a design choice.

The patient pathway is, by definition, a fundamentally patient-centric way of looking at healthcare. It is one particular lens through which to view healthcare processes, and one specifically designed to centre the processes from the patient's perspective, suggesting that this one thread should be focused on. Whether the object-centric perspective is useful for patient pathways or not, it is clear that the increasing popularity of object-centric paradigm in PM research is naturally leading towards graph-based process models, and the increasing incorporation of methods that resemble those used by knowledge graphs.

2.6 *Unified Modelling Language*

Chapter 5 of this thesis makes extensive use of the Unified Modelling Language (UML) standard, a language designed to provide a way of describing the architecture and design of systems (OMG 2017; Rumbaugh et al. 1999). Whilst it is best known for its applications in software engineering, it is in principle general-purpose. As well as providing a system for creating

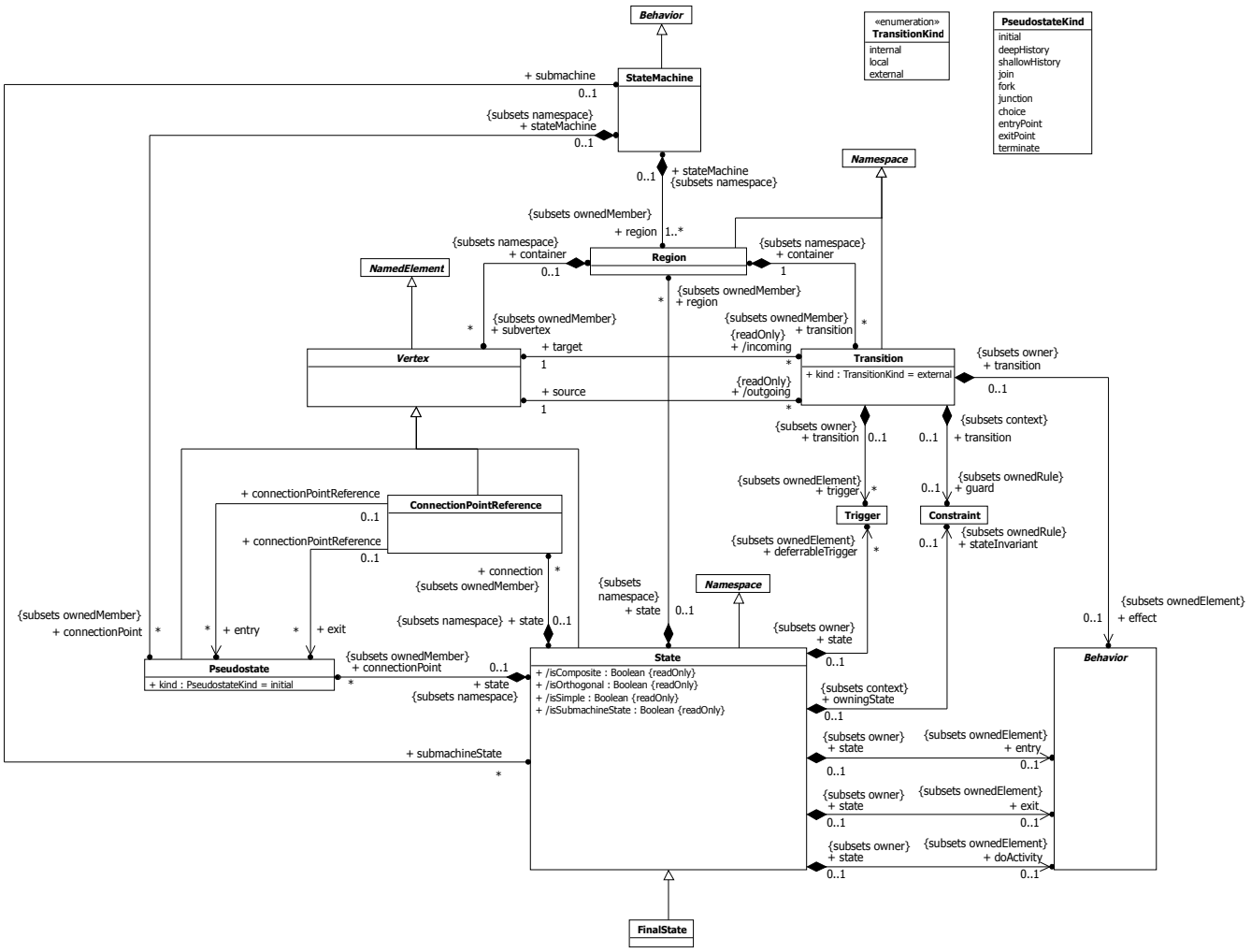

Figure 2.1. The UML state machine formalism, described as a UML class diagram (OMG 2017)

activity diagrams, which are used to describe software systems, the standard also encompasses several other notations, including the *state machine*. The UML state machine focuses on the state of a system at a point in time, rather than the flow of control. Based closely on Harel statecharts (Harel 1987), it extends the traditional notion of a finite-state automaton (Hopcroft & Ullman 1979) with a handful of useful features such as nested states and orthogonal regions UML state machines are used in Chapter 5.3; their fundamentals are therefore introduced here for reference.

The full state machine standard is relatively complex; Figure 2.1 describes it as a UML class diagram. Here, each box (*class*) represents a type of entity, and the box's contents describe that entity's attributes. Lines describe relationships between entities, which can be a simple association ($\wedge$), or describe an inheritance ($\triangle$) or composition ($\blacklozenge$) relation. Numbers indicate the cardinality of relationships, for example `0..1` indicates that a relationship occurs between 0 and 1 times, with * indicating any number.

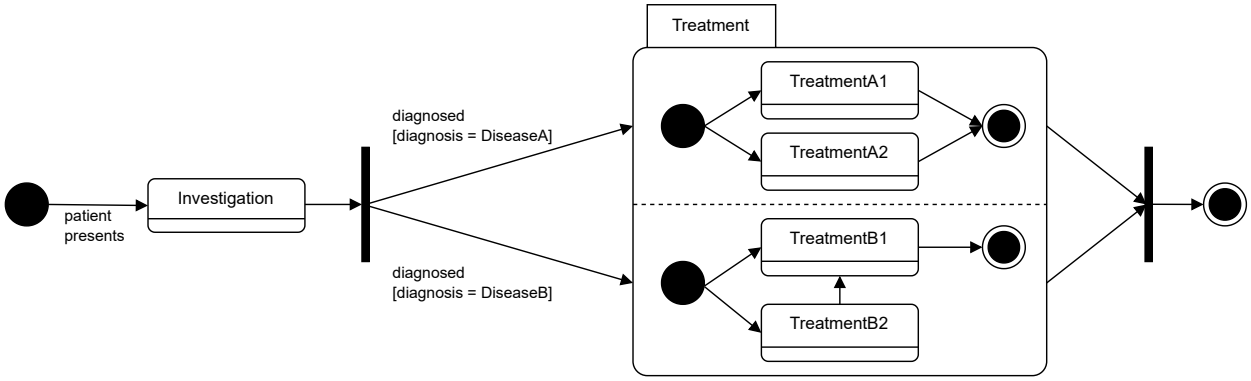

Figure 2.2. An example UML state machine

A UML state machine is comprised of *regions* — sections of the graph that can execute concurrently — each containing a set of *vertices* interconnected by *transitions*. A *vertex* can be the source and/or target of any number of *transitions*. Vertices form an abstract class, its specific semantics largely depending on its type: *states* are stable, meaning execution remains there for some time, whilst *pseudostates* are transitive, with execution passing through them automatically.

The most important type of vertex is the *state* (⊟), which models a situation in the execution of a state machine in which some condition holds. States can be *simple*, containing no internal vertices or transitions; *composite*, containing at least one region; or *submachine* states, containing an entire state machine. States can have associated entry and exit behaviours which are executed when transition occur. They can also have an associated *behaviour*, which commences when the state is entered (but after the entry activity completes) until either it completes or the state is exited. A *final state* (⊙) is a particular type of state that indicates that the enclosing region has completed execution. *Pseudostates* are a family of vertices which are largely transient: that is, execution usually passes through them. This is mostly used to represent control flow, such as join, fork, or choice points. An *initial* pseudostate (•) marks the entry point for an entire region; it is the source for exactly one transition, and there is only one per region.

A *transition* (→) is a directed arc between a source and a destination vertex, which can be the same vertex. The exact semantics of a transition depends on its relationship to its source vertex. An *external* transition exits the source vertex, and therefore executes any associated exit behaviour for that state. A *local* transition does not exit its containing stage, and therefore executes no exit behaviour — these can only exist within composite states. *internal* transitions are local transitions with the same source and target, so no exit or entry behaviours are executed. The transitions associated with a

state can be further associated with *triggers* and *constraints*, and a transition can be either instantaneous or have an associated duration.

Execution is triggered by the occurrence of *events*: an execution of a state machine is a set of valid path traversals through one or more regions' graphs, triggered by event occurrences that match the triggers and constraints in those graphs. Figure 2.2 shows an example of a UML state machine. A patient presenting themselves is an event which triggers a transition to the investigation state; receiving a diagnosis (a trigger) moves the patient into one of two parallel regions via a split/join pseudostate (|) depending on the disease they are diagnosed with (a constraint). A diagnosis of disease A means they are treated with either treatment A1 or treatment A2; a diagnosis of disease B means they are treated either with treatment B1, or treatment B2 followed by B1.

2.7 *Summary*

In summary, the idea of the patient pathway is widely used in modern healthcare, but the tools to share, analyse and evaluate them are lacking. Pathways are analysed based on binary groups, when the reality is that a patient's relationship to a pathway is a continuous scale, requiring some sort of similarity metric for informative analysis to take place. Quantifying the distance between a patient's true path and their recommended pathway could lead to more informative evaluations of these pathways, but comparison is difficult in the absence of any formal representation of pathways. They are written and shared as free text and ad hoc diagrams, and whilst attempts have been made to formalise them into languages, these are mostly designed for point-of-care decision support rather than retrospective analysis, and none have been widely adopted for this purpose.

A number of methods have been proposed for analysing patient pathways. PM is probably the most widely used and well-developed, but the methods by their general nature tend to struggle with health data, and specialised methods for health data are still in their relative infancy. However, in some circumstances, such methods may be a useful approach to producing summary or "average" pathway diagrams for different groups. The electronic phenotyping literature provides a rich source of methods for grouping patients in terms of diagnoses and symptoms, meaning that these techniques could be adapted to focus on the interventions and procedures under clinician control by viewing pathway analysis as a sort of "pathway phenotyping". Across all the different approaches to pathways however, a clear running

thread was the advantage of incorporating relevant domain knowledge into general techniques.

Knowledge graph methods are increasingly popular, and have a number of properties which make them well-suited to representing and analysing complex healthcare data. In particular, their structures supports both logical reasoning and ML approaches to analysis, and they support the m erging of both observational data and domain knowledge. They have been applied to a wide range of biomedical problems, but as of yet few attempts have been made to use KG formalisms to solve pathway problems. Through either structural knowledge (logical relations between concepts) or learned values (embedded distances between concepts), graph reasoning could help to encode concepts such as semantic similarity into patient pathway models, allowing for more informative representations of pathways.

3 *The structure and interpretation of patient pathways*

Pathways represent the journeys taken by patients through the healthcare system. Where recommended pathways are specified, they are guidelines rather than rules, and are subject to interpretation and decision-making by clinicians at every point. Data from EHRs is a useful source of information on pathways, but extracting meaningful insights is difficult due to the sheer complexity and heterogeneity in pathways. The existence of pathways raises a number of questions around compliance and effectiveness that are difficult to meaningfully measure, but before these can be answered the realities of the data being worked with need to be considered.

This chapter therefore considers the relationship between pathways and the actual data that describes them. Section 3.1 covers the necessary preliminaries, establishing a precise definition of a pathway that is suitable for the purposes of this thesis, and providing an overview of the structure of modern electronic health record data along with the various standards and terminologies that are used. I then consider how patient pathway data can be interpreted from two distinct perspectives.

I examine the patient's *context*, i.e. the wider factors that might affect their treatment pathway. By summarising the real-life pathways of a population of colon cancer patients, and then breaking down the cohort according to various factors, I examine the extent to which factors such as age, deprivation, and other comorbidities affect the shape of their pathways (Section 3.2).

I then consider the event data itself, and attempt to establish rules and logic behind which events are included and which events are ignored in analysis. I outline an approach that uses structured knowledge from ontologies to automatically infer relationships between diagnoses and procedures, and provide empirical evaluation of its effectiveness compared to conventional methods (Section 3.3). These two contrasting approaches — one considering human factors and context, the other considering logical rules and the structure of data — together help us to understand exactly what knowledge is encoded in healthcare data, and what is not.

3.1 *Contextual and structural knowledge in pathways*

This section covers two necessary preliminaries for thinking about how patient data is interpreted. Firstly, I discuss different definitions of patient pathways, and arrive on one suitable for our purposes. Secondly, I outline how EHR data is structured and coded, and introduce the major ontologies used in UK EHR data.

3.1.1 *Defining pathways*

The terminology surrounding pathways can be confusing, with different source referring, often interchangeably, to "pathways", "guidelines", "protocols", and "care maps" amongst others. De Bleser et al. (2006) analysed three years of publications on the topic and were unable to agree on a single definition. Alternatively, pathways can be defined by what they are *not*; this is perhaps easier since many definitions make a point of drawing a distinction between *pathways* and *guidelines*. A clinical *guideline* is a statement of recommendations for optimising patient care, backed up by systematic reviews of the evidence on a particular topic. In the UK, the most notable examples are the guidelines produced by the National Institute for Health and Care Excellence (NICE). A clinical *pathway*, meanwhile, is a more localised set of recommendations that describe the specific processes used by a particular healthcare provider. These are typically less concerned with specific physician-patient interactions, and instead describe operational and logistical considerations, ensuring effective flow of information between all clinicians involved in a patient's care. (Panteli et al. 2019; Rotter et al. 2019).

Possibly the most comprehensive definition comes from Kinsman et al. (2010). Attempting to unify the existing terminology, they define a clinical pathway as a "structured multidisciplinary plan of care" that meets any three out of four conditions:

1. it channels the *translation* of guidelines into more local structures;
2. it *details the steps* in a course of treatment as an algorithm or protocol;
3. it describes either *timeframes* or *criteria-based progression*;
4. it aims to standardise care for a *specified* clinical problem, procedure or episode in a *specified* population.

By this definition, clinical pathways act as a bridge between guidelines — what *should* happen — and real care processes — what *does* happen — that translate and tailor guidelines to the specific circumstances and environment of a care provider.

The *system of concepts to support continuity of care*, or *ContSys* for short, is an international standard that defines several key concepts surrounding

the continuity of care (ISO 2015; Oughtibridge 2019). This therefore provides a structured way of describing types of healthcare event, and how they relate to each other. In particular, it contains five concepts that relate closely to clinical pathways:

- a *clinical guideline* is a set of statements to assist clinicians with decision-making
- a *clinical pathway* represents the best-practice workflow for a particular diagnosis, defined on a national level
- a *core care plan* refines the clinical pathway, and is defined by a particular healthcare provider
- a *patient care plan* further refines the core care plan to tailor to the needs of an individual patient
- and a *patient journey*, which is the actual timeline of treatment for a patient as it occured.

These concepts describe a spectrum of care processes: at one end, a set of abstract guidelines that provide broad guidance for certain topics, and at the other, the patient's journey as it happened, from which retrospective datasets are derived and researchers see (Figure 3.1). At each level, guidelines are interpreted by different users to meet increasingly specific requirements. As these requirements become more specific, the set of possible values becomes a subset of those possible at the preceding step — with the exception of the true patient journey, where unforeseen external events are possible. Much like the definitions previously discussed, the ContSys model places a clinical pathway somewhere in the middle of the scale: it implements and specifies clinical guidelines, and it results in a particular patient journey.

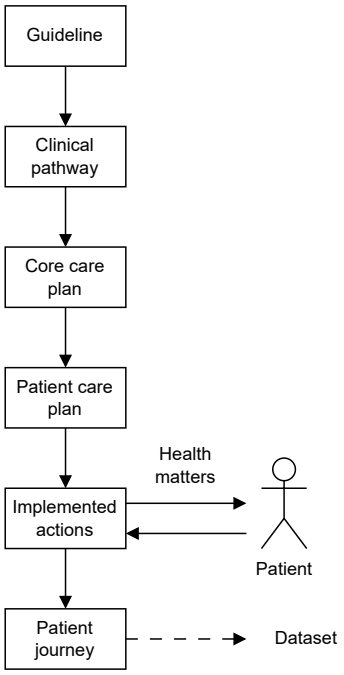

Figure 3.1. The layers of a patient pathway, according to the ContSys model

The ContSys standard is also a useful reference point for patient pathways because it draws a distinction between *direct* events, which represent interactions between the patient and the clinician, and *indirect* events, which represent activities without the patient present, such as administrative procedures. Rojas et al. (2016) draw a similar distinction, contrasting data from administrative systems with data from clinical systems, and analysis of treatment processes with organisational processes.

This differentiation between direct and indirect events is a useful way of describing data, although it may be overly simplistic in some situations, since an indirect or administrative event can still impact a patient. Administrative procedures such as referrals, for example, can affect how long it takes for a patient to receive their diagnosis or treatment, and such delays can have a measurable effect on patient outcomes (Neal et al. 2015). Indirect events can also be useful proxies where data on direct events is not available, albeit with

caveats: for example, prescribing data can provide an estimate of medication usage, but it cannot provide any guarantee that the medication was actually administered.

These definitions and standards collectively demonstrate that clinical pathways are defined by their position on a spectrum of evidence-based care documents and their role in translating guidelines into clinical practice, rather than by adherence to any specific format. For the purposes of this work, I therefore use "clinical guidelines" (CGs) to refer to high-level descriptions of broad recommendations, often set on a national scale, and "clinical pathways" (CPs) to refer to "true" CPs according to Kinsman et al.'s definition, i.e. a concrete, local implementation of a guideline. Additionally, I use *trace* to mean the actual sequence of events a patient experiences, as visible in retrospective datasets.

3.1.2 *Ontologies, terminologies, and clinical coding*

In electronic health record (EHR) systems, and datasets derived from them, the key concepts that make up a patient's history are represented by *clinical codes*, alphanumeric identifiers that represent particular concepts. In principle, these coding systems exist to standardise the reporting of events, and to allow medical records to be converted into terms that can be analysed and compared across different computer systems, hospitals, and countries. In practice, healthcare systems encode their data in a wide range of classifications, terminologies, and ontologies that vary between, and even within, countries and healthcare systems (Haendel et al. 2018). In the United Kingdom, clinical data is largely recorded according to three major terminologies.

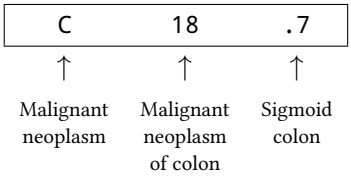

Figure 3.2. An ICD-10 code and its structure

ICD-10 (the International Classification of Diseases, Version 10) is an international standard maintained by the World Health Organization. It describes diseases, signs, symptoms and findings, and is used worldwide to enable comparison of mortality rates and causes. In older versions such as ICD-9, codes also exist to represent procedures and treatments (Volume 3 or "ICD9Proc"); the more recent ICD-10 only covers diagnoses, but the USA maintains a local extension, the ICD-10 Procedure Coding System (ICD-10 PCS), for encoding procedures. In the future, ICD-11 will expand this with new classifications of signs and symptoms, and semantic features as well as closer alignment with other standards such as SNOMED, however it

OPCS-4 (the OPCS Classification of Interventions and Procedures, Version 4) is the coding system for medical procedures used by NHS hospitals in the UK, derived from the much older Office of Population Censuses and Surveys' Classification of Surgical Operations. 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.

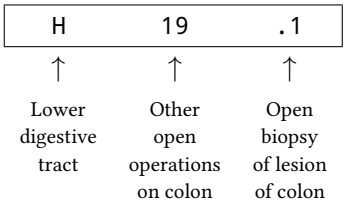

Figure 3.3. An OPCS-4 code and its structure

SNOMED CT is an international ontology containing over 350,000 medical concepts. Its structure describes semantic relationships and polyhierarchies amongst over 350,000 entities, allowing the relationships between concepts to be described in detail, rather than simple parent-child relations. This semantically rich structure makes SNOMED a true ontology, rather than a classification or terminology, and means that SNOMED concepts and relations can be reasoned over using logical rules to make inferences about healthcare concepts. The SNOMED standard specifies a language for this exact purpose: the Expression Constraint Language (ECL; SNOMED International 2022).

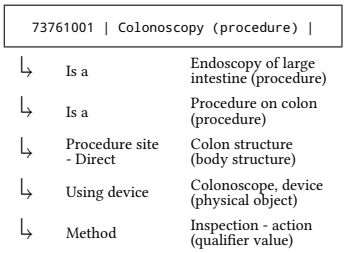

Figure 3.4. A SNOMED CT concept and its immediate relations

These standards are designed and used for different purposes, rather than to replace one another. ICD and OPCS are strictly speaking *statistical classifications*: a type of terminology in which concepts must be mutually exclusive and arranged mono-hierarchically (every concept having exactly one parent). This rigid structure is deliberately designed to make statistical reporting as easy and unambiguous as possible (Haendel et al. 2018). Whilst SNOMED has been a mandated standard in the UK since 2020, datasets are still widely published and shared with ICD diagnoses and OPCS procedures because of these benefits for reporting, epidemiology, and reimbursement (NHS England 2020, 2023c, 2024).

3.2 *Context and interpretation in pathway analysis*

Process mining — a field of research introduced in detail in Section 2.4.4 — is often described as the study of, or a set of techniques for the analysis of, historic event data (Munoz-Gama et al. 2022; van der Aalst 2012, 2016). Real-world processes do not exist in a closed system: they are designed, implemented by, and interact with humans, and in no area is this more important than in healthcare. A treatment process is more than a sequence of events or a conveyer belt of patients — it is naturally non-deterministic, and shaped by human decisions, priorities, and biases. Before the data and the processes themselves can be explored, the context in which the data was produced must first be considered. This section therefore investigates the extent to which treatment pathways are influenced by contextual factors.

3.2.1 *Background*

The bulk of data in a patient's health record is represented in the form of clinical codes, which represent healthcare concepts and are defined according to a standardised scheme, but the interpretation of these codes does not depend solely on their meaning as assigned by the coding system — it also depends on the context and purpose of the coding process.

This affects EHR datasets on multiple levels. Firstly, patient records are not designed to be used for retrospective studies: they are, before anything else, a practical tool to be used at the point of care and as an *aide-mémoire* for clinicians (Goldacre & J. Morley 2022). The care record is not, therefore, simply a factual record of what happened to a patient and when: it might aspire to this, but it is also a reflection of human processes and priorities. Entries are influenced by the clinician's decision on what was important to record, by the standards and practices of the particular hospital, and sometimes by design elements such as the user interface (Madandola et al. 2024).

Secondly, clinical coding — the process by which clinician-generated records are translated into standardised terminologies — exists for reimbursement and resource planning purposes. Codes are typically assigned by clinical coders based on the notes available to them, although some must be assigned by clinicians, especially those requiring interpretation of numerical test results (Nouraei et al. 2016). Where a clinical coder does assign codes, they are reliant on the quality of documentation by the original clinician. The resulting structured data is *useful* for epidemiology and public health research, and is widely used to perform research, but is not what the record was originally designed for. They are also not the whole story: a clinician's decision making at the point of care is more likely to have been based on, for example, the specific value of a blood test result[1], rather than the existence of a diagnosis code that may not have been assigned until later.

1 for example, in diabetic ketoacidosis, great attention is paid to measuring blood ketones (Kilpatrick et al. 2022).

In addition to the layers of interpretation involved in producing a healthcare record, the care process itself can also be influenced by underlying factors. Perhaps the most notorious pattern found in health records is the *weekend effect*, which describes the supposed tendency of hospitals to exhibit higher mortality rates at weekends compared to weekdays. This has often been blamed on lower staffing levels at weekends (Freemantle et al. 2015; McKee 2016). However, a number of studies have argued that patients who present at weekends have different characteristics to those who present on weekdays. They are less likely to have been referred by their general practitioner (GP) and tend to be more acutely, rather than chronically, ill

(Bion et al. 2021); there is typically a reduced availability of alternatives to hospital (Walker et al. 2017); and in some cases the threshold for admission is higher at weekends, meaning that only the more severe cases appear in the statistics (Meacock et al. 2017). This example illustrates the existence of contextual factors which are not immediately visible in the data itself, but which nonetheless influence it. Often, these are the result of human processes and practices: for example, in some healthcare settings it is common to assign a patient a diagnostic code to indicate a suspected diagnosis, meaning it might appear in the patient's record even if it was later ruled out (Atolagbe et al. 2024). Fawcett et al. (2019) try to quantify this, by comparing the number of endiocarditis diagnostic codes to the number of verified cases. They bridge the gap between these two different numbers by applying domain knowledge and formulating a set of rules: removing codes that are known to be unreliable, removing very short hospital admissions without death, and removing readmissions within 30 days. A practical understanding of the subject area and an understanding of the people, processes and organisations that generate the data is required for results to be clinically meaningful.

In statistics, Simpson's paradox describes the phenomenon whereby trends that exist within particular subgroups can be hidden, or even completely reversed when the entire population is analysed (Simpson 1951, Figure 3.5). Simple or population-wide analysis is rarely enough when it comes to understanding data, and expert subject-matter knowledge needs to guide any analysis, and consider the different groups that make up a population (Hernán et al. 2011). In pathway terms, this means that pathways may look very different for different populations, and that contextual factors needs to be considered.

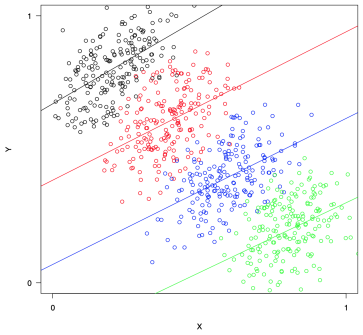

Figure 3.5. An example of Simpson's paradox: X and Y appear to be negatively correlated, but within each subgroup the correlation is actually positive (Rücker & Schumacher 2008)

Therefore, EHR data is rarely a pure and objective record of treatment, but subject to two layers of interpretation at the very minimum: by the clinician, in deciding what to record and in how much detail; and by the clinical coder, in how they choose to translate this, as well as potentially many more layers, including design choices made in the coding system. Whilst electronic health records summarise the events that occur in a healthcare process, they are not an objective record, nor do they tell the full story. In addition to these issues which affect how healthcare data can be interpreted, it is also vital to consider a patients' record beyond the events which make up their treatment history. Patient-specific context such as age, gender, ethnicity, social background and more will all influence the relationship between a patient, a disease, and their treatment pathway.

Typically, attempts to reduce model complexity in the process mining literature often focus on the data itself: attempting to cluster into a mean-

ingful number of groups, or filter irrelevant ones out (Aspland et al. 2021; Munoz-Gama et al. 2022). The examples discussed here prove that many organisational and human factors affect the information encoded in a health record, and that an effective study of patient pathways needs to consider these at a deep level. This section therefore provides an investigation into the importance of contextual data, and investigate how these principles apply to patient pathways.

3.2.2 *Methods*

To investigate the extent to which contextual factors affect understanding of patient pathways, I analyse the patient pathways in a real-world dataset, and stratify the population by several characteristics to see how they affect the conclusions. The dataset consisted of anonymised patient records for patients treated for colorectal cancer between December 2012 and February 2024 inclusive by Oxford University Hospitals NHS Foundation Trust (OUH), collected by the National Institute for Health and Care Research (NIHR) Health Informatics Collaborative (HIC) programme (Tamm et al. 2022).[2] The cohort was defined as any patient with a diagnosis of colon cancer (ICD-10 code `C18`) in the hospital's cancer patient database ($n = 1,571$). Patients whose date of diagnosis was less than one year before the dataset's upper date limit were discarded, to ensure that sufficient follow-up data on patients' treatments was available. This created a cohort of 1,417 patients.

2 The collection of this data received ethical approval from the NHS Health Research Authority (East Midlands - Derby Research Ethics Committee, `21/EM/0028`)

Given the available input data (the EHR's inpatient and outpatient procedures table), each patient's history is extracted as a chronologically ordered sequence of events, where each event consists of an OPCS code and a timestamp. This list is then filtered to contain only events relevant to colorectal cancer. Relevant OPCS codes were identified using an established list of codes from the national Colorectal Cancer Data Repository (CORECT-R) project (Downing et al. 2021). This major research study maintains a data dictionary listing 224 procedure codes defined by clinicians as relevant to colorectal cancer, and maps them into several broad categories (*CORECT-R Data Coding* 2020). This codelist did not, however, cover several important categories including chemotherapy and radiotherapy; it was therefore enhanced with these extra codes before use in consultation with domain experts.

The events in these filtered event logs are grouped into a smaller number of categories, based on those in the CORECT standard, and the filtered and abstracted log is used as input to generate directly-follows graphs (DFGs), diagrams which visually summarise the frequency with which particular events are followed by other events. In these graphs (Figures 3.6–3.9), each

node is labelled with the *relative case frequency*, i.e. the proportion of all patients who experienced that event, whilst edges are directed and labelled with the the *relative antecedent frequency*, i.e. the proportion of instances of the source event that were followed by the destination event.

3.2.3 *Results and discussion*

Figure 3.6 shows the overall summary of treatment pathways for all patients in the cohort. A majority of patients (61.96%) received some form of resection, with 55.33% receiving major resection and 27.73% receiving minor resection.[3] 40.16% of patients underwent chemotherapy, and 6.56% underwent radiotherapy. 18.98% of patients had no recorded treatment at all.

Official statistics indicate that 63% of tumours were treated with resection in the Thames Valley region between 2013 and 2021, 33% received chemotherapy, and 5% received radiotherapy. 24.5% received "other care" (NDRS 2021). These figures — allowing for minor differences in cohort selection and codelists, the different study period, and the fact that the official statistics are counted by tumour rather than by individual — are therefore mostly in line with what would be expected.

3 These two figures total more than 61.96% because patients can, and do, undergo multiple surgeries.

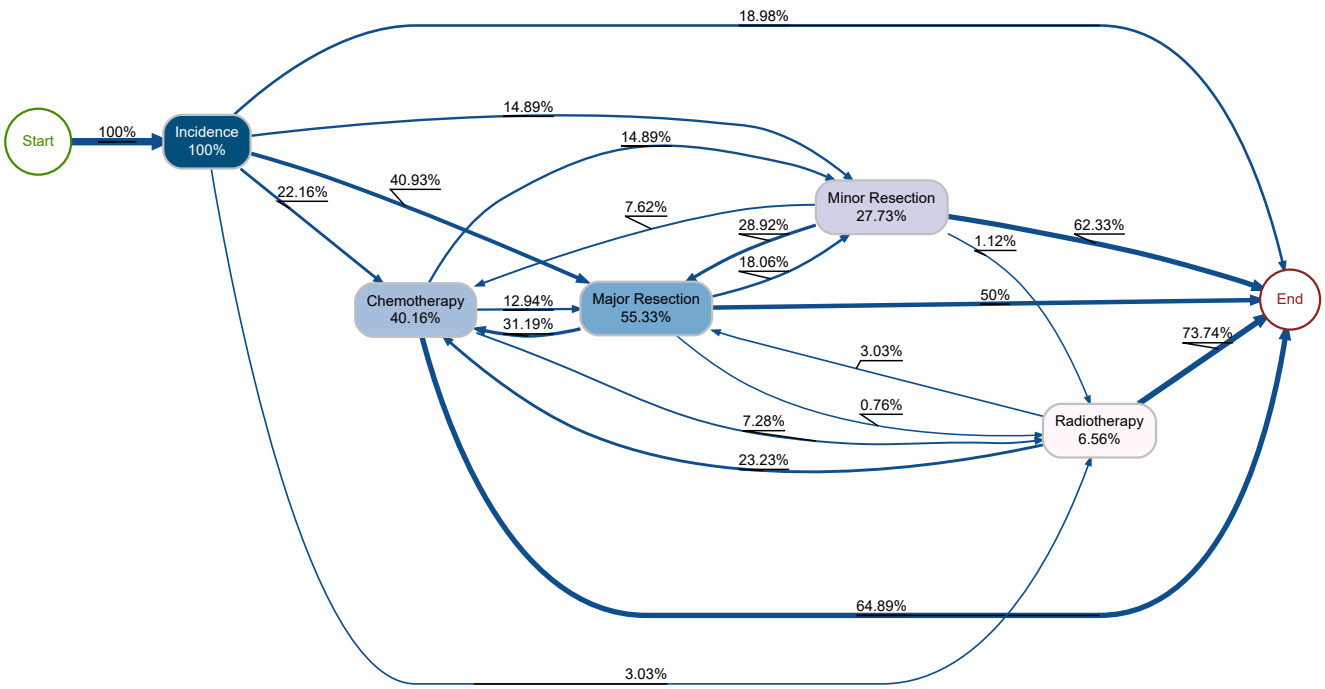

Figure 3.6. Treatment pathways for all colon cancer patients

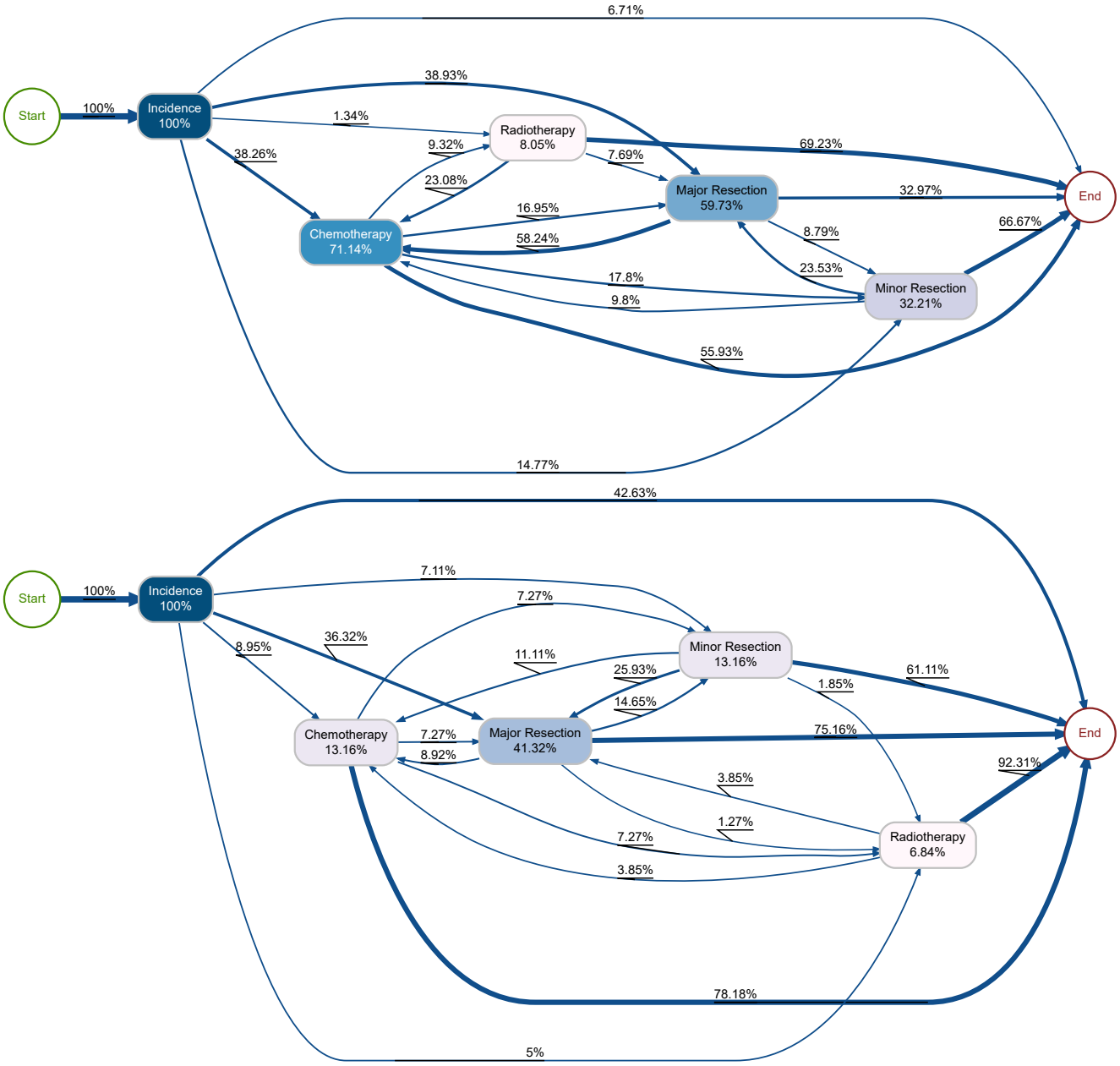

Figure 3.7. Treatment pathways for colon cancer patients aged 50–59 (top) and 80+ (bottom)

Stratification by age

Figure 3.7 shows the pathways as they appear in two different age groups. In the 50–59 age group, a large majority of patients (71.14%) underwent chemotherapy, whereas only 13.16% in the 80+ age group did. There were also significant reductions in the proportions of patients receiving resections, and a smaller reduction in radiotherapy. Furthermore, 42.63% of patients aged 80+ received no known treatment at all, compared to only 6.71% of those aged 50–59.

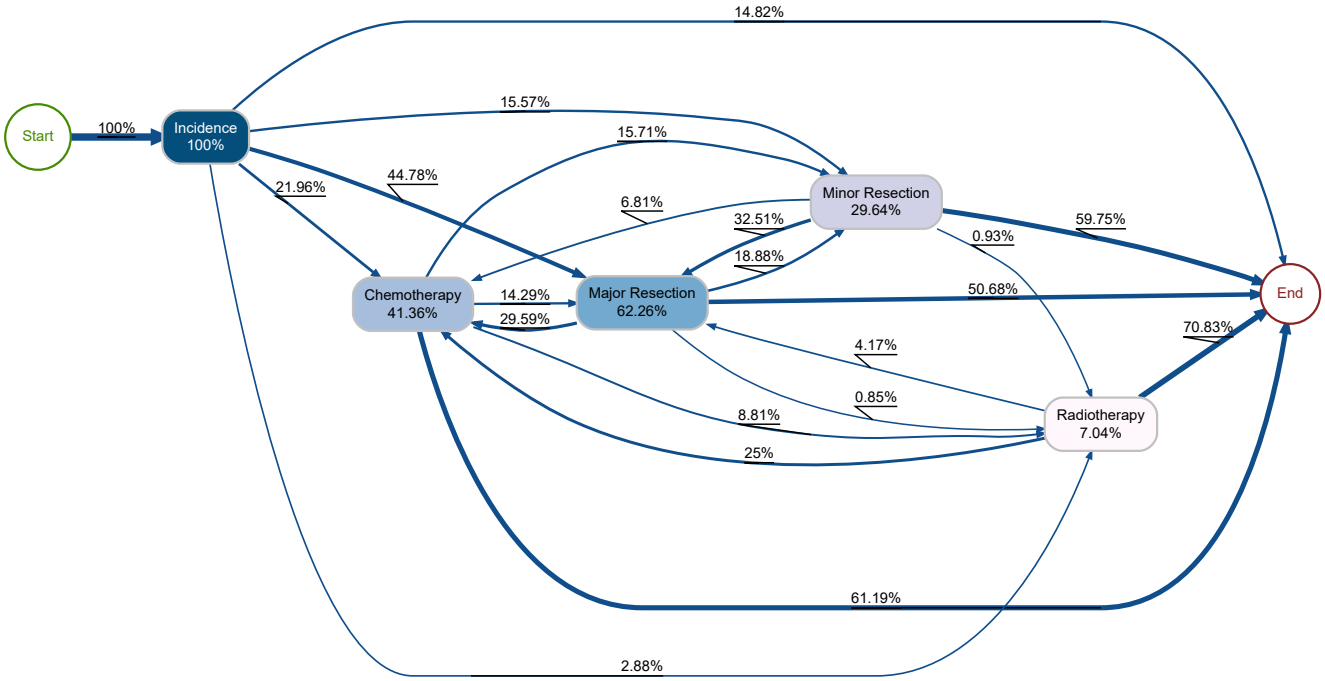

Figure 3.8. Treatment pathways for colon cancer patients with a Charlson comorbidity score of zero

This pattern aligns with existing evidence, which suggests that the range of treatment options narrows with age. In part, this reflects clinician's decision-making: increased age typically correlates with comorbidity, meaning the risk of side-effects is increased, and more aggressive treatment regimens become riskier. However, patients have the right to be consulted on and involved in all decisions about their treatment (*NHS Constitution for England* 2012), and research indicates that younger patients typically prioritise treatment based on increased life expectancy, whilst older patients tend to value quality of life (Shrestha et al. 2019). The fact that treatment pathways vary so noticeably by age highlights the fact that clinicians are not the only decision-makers in a treatment process: the voices of patients are also valuable, and individual preferences and values affect the patient pathway.

Stratification by comorbidity

The Charlson comorbidity index is a widely used measure of mortality risk. Patients are scored based on the presence of seventeen different comorbid conditions weighted by severity, and points are added for increased age (Charlson et al. 1987). Figure 3.8 shows the pathways of patients with the lowest possible Charlson score of zero, indicating none of the significant comorbidities. The proportion of patients with no known treatment at all is noticeably lower in this group (14.82%) compared to the whole population

(18.98%). This non-treatment rate increases to 24% and 47% in patients with scores of 1 and 2 respectively.

Conversely, the major resection rate decreases from a baseline 62% in those patients with a zero score to 45% in 1, then to 25% in 2. Increased comorbidity typically entails increased frailty, and therefore a reduction in a patient's tolerance or willingness to undergo more intensive treatments. National data puts the proportion of patients with Charlson scores of zero receiving "other" treatment at 22% (NDRS 2021), which perhaps indicates localised difference in practice.

Stratification by deprivation

The index of multiple deprivation (IMD) is a measure of socio-economic deprivation in England, calculated by dividing the country into 32,844 areas and ranking them according to income, employment, education, health, crime, barriers to housing and services, and living environment (Ministry of Housing, Communities & Local Government 2019). Figure 3.9 shows a summary of pathways in the most and least deprived areas, defined as the first and fifth IMD quintiles respectively. Similar numbers of patients received no known treatment (16.67% in the most deprived areas, 17.45% in the least deprived areas). whilst patients in less deprived areas were slightly more likely to undergo major resection (52.06% vs 55.61%) and minor resection (22.92% vs 29.75%), but significantly less likely to experience chemotherapy (52.08% vs 40.50%) and radiotherapy (10.42% vs 5.51%).

Previous research has indicated that patients in more deprived areas tend to have lower rates of survival (Møller et al. 2012; Syriopoulou et al. 2019). A number of reasons have been suggested for this, in particular that patients tend to present at a later stage of disease (Lejeune et al. 2010). Mapping patient pathways allows us to paint a fuller pictures of these disparities, and reveals that these pre-existing factors also lead to differences in the treatment itself. Further analysis is required, however, to establish the exact mechanism of action: whether later-stage presentation entirely explains these differences in treatment, or whether other factors such as lack of access to treatment options, travel distances to healthcare facilities, or clinician bias might also contribute.

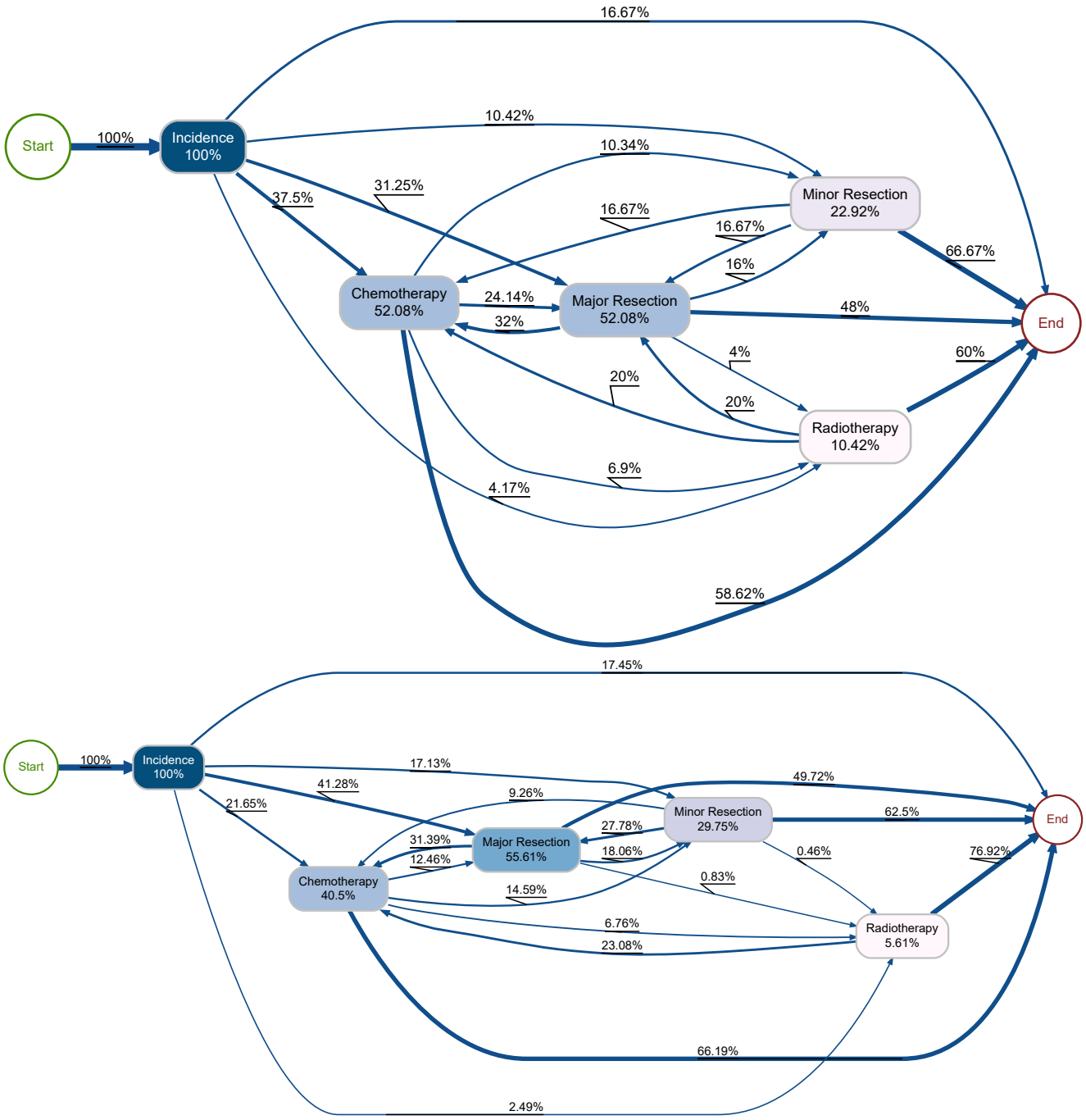

Figure 3.9. Treatment pathways for colon cancer patients in the most (top) and least (bottom) deprived quintile

3.2.4 *Summary*

Each of Figures 3.7–3.9 shows that patient journeys look significantly different depending on which part of the population is being studied. Moreover, these differences are often explainable and can be related to well-observed differences in the medical literature. This demonstrates that event data is, by itself, not enough to deeply understand treatment processes: an examination

of pathways needs to consider contextual factors. Pathways are influenced by a myriad of variables. These include observations that exist in the dataset but are not strictly event data, such as a patient's demographic characteristics, but it can also include factors not in the data, such as the processes and policies used by a particular healthcare provider, or the preferences of the patient or clinician.

Contextual information is not a silver bullet that instantly explains all variation in processes — in particular, the example of deprivation highlights how much research is still to be done — but it does confirm that analyses of whole populations are of limited use, and that any research into patient pathways needs to carefully consider the characteristics and idiosyncrasies of the particular population and disease.

3.3 *Defining pathways with ontologies*

As well as the contextual factors that affect the events coded in an EHR, a good analysis needs to incorporate a thorough understanding of the data itself. Healthcare processes are complicated because they are characterised by incomplete and noisy signals, high levels of process variation, and multitudes of exceptional behaviours that should ideally be captured rather than disregarded. Preparing this data for analysis is therefore a significant and time-consuming step, where modelling decisions and assumptions can have a significant impact on the eventual results. Typically, when determining which clinical events to include or exclude for analysis, the options are to either manually curate a list of events of interest, or to use statistical factors to distinguish between common events and outliers, both of which come with advantages and drawbacks.

In this section, I propose a third approach to preparing raw healthcare data for process analysis: an ontological approach. Given that healthcare data is typically encoded in standardised terminologies, and that these terminologies are mappable to ontologies rich with semantics, this structured knowledge could be used to facilitate the process of event log generation. By specifying domain-informed constraints, plausible relationships between diseases and events can be deduced, helping to reduce a large dataset of events to a smaller subset of relevant ones, making for a more focused and informative model. This sections describes this proposed approach, and evaluates it on a dataset of electronic health records.

3.3.1 *Background*

Determining which events in a patient's history should be included or excluded in analysis is a fundamental stage of data analysis, which has a significant impact on results, but it is one that is not always described in detail in process mining publications. De Roock and Martin's literature review (2022) finds that only 22% of process mining papers analysed mentioned data pre-processing, and only 13% mentioned data filtering. Emamjome et al. (2020) describe similar results, observing that 72% of papers either used relatively naive approaches to preprocessing — that is, not considering data quality issues and not relating data pre-processing to their research question — or did not describe it at all.

The problem of filtering a patients' history has, however, been studied before in the process mining literature: where there is a high level of variability in possible traces, the most common approach is to filter at the event and the trace level (Marin-Castro & Tello-Leal 2021). However, *filtering* is often framed as a problem of removing noise and logging mistakes.[4] Determining whether a recorded event actually happened, or happened as described, is undoubtedly a question of interest when studying healthcare records, but there is another step that must come first. Given that a patient will typically have a long history with a large number of possible events, the first question should be whether a particular event is even relevant to the disease or pathway being studied.

4 for example, Marin-Castro & Tello-Leal (2021) describe filtering as "determin[ing] the likelihood of the occurrence of events or traces based on its surrounding behavior"

An easy and intuitive heuristic that can quickly simplify a large event log is to only examine the most common events in the dataset. However, infrequent events can still be very relevant to a process and have an effect on it; conversely, frequently occurring events can be completely unrelated to the process. This notion that infrequent behaviour should be captured and analysed rather than discarded as noise is a key idea in the current thinking around healthcare process mining: Munoz-Gama et al. (2022) note that "researchers and practitioners must go beyond simply filtering out infrequent behaviour from the event log". Tax et al. (2019) demonstrate that filtering activities based on frequency alone does not solve the problems of what they term "chaotic activities" — events that occur independently of the state of the process — and ultimately affects the quality of the final process model.

The second simple approach is to hand-curate a list of events to include. This is the norm in healthcare and epidemiology research: researchers studying health records typically create a *codelist*, a list of concepts in the dataset's terminology, which enumerates which exposures and outcomes are being

investigated (Williams et al. 2017). The development of these lists is a long process, requiring a clear definition of the clinical feature of interest, a short-list of the potential codes to include, and an iterative process of expert review (Watson et al. 2017). It is therefore naturally very time-consuming, and must be repeated for different datasets using different terminologies. However, the advantage is that it produces a high-quality codelist, rigorously defined and approved by experts in the field.

Increasingly, attempts have been made to collect codelists in online repositories, to establish standardised definitions or *phenotypes* for particular diseases and to improve the reproducibility research. However, these libraries are predominantly focused around lists of diseases rather than procedures, with their codelists largely consisting of diagnostic codes (*HDRUK Phenotype Library* 2023; *OpenCodelists* 2023). It has been noted that relatively small variations in diagnostic codelists can lead to noticeable differences in the measured outcomes (Makadia et al. 2023): it therefore follows that similar changes in a procedural codelist would lead to variations in process models, and greater attention therefore needs to be paid to them. Given that process mining involves a process of curating procedural codes to extract from a healthcare record, which is analogous to the curation of diagnoses to select a cohort, the field of healthcare process mining might benefit from an equivalent repository of procedure- and event-centric codelists, to similarly establish common phenotypes of disease pathways and facilitate reproducible research.

Cremerius et al. (2023) propose a standard approach to generating event logs from healthcare datasets for process mining. In this approach, event filtering involves human curation, with researchers choosing procedures in a terminology based on those described in relevant medical guidelines. However, this approach is limited by the typically vague language of medical 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 (NICE 2020). Guidelines will also differ by location, making comparing results difficult. There is therefore a need to explore this event selection stage in deeper detail, and establish clear and consistent rules for what is and is not included.

The principle of using semantics and reasoning to assist in the preparation of codelists has been explored before: Elkheder et al. (2023) demonstrate how reasoning over SNOMED relationships can be produce diagnostic codelists, and that these codelists produce similar results to their handcrafted

equivalents when used to select patients from a 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 each other.

The use of reasoning to enhance process mining has also been explored: Alves de Medeiros and van der Aalst (2009) describe *semantic process mining* as the explicit relation or annotation of elements in an event log with the concepts they represent, thereby making it possible to automatically reason or infer other relationships between them, and even suggest that such ontologies might be useful in the log cleaning process. Other approaches have used reasoning for *abstraction*, grouping group granular events into categories for analysis, and for *aggregation*, merging frequent event sequences into single events based on rules (Leonardi et al. 2019; Remy et al. 2020). Work has also highlighted the apparent gap between the 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 (Klessascheck et al. 2021).

An automated — or at least semi-automated — codelist curation process has been attempted before: for example, Watson et al.'s framework (2017) selects codes based on whether their full text descriptions contain particular search terms. The authors recommend that this step should err on the side of caution, retaining uncertain codes unless there exists a clear consensus to reject them. Clinical datasets, however, already contain useful semantics embedded in them by virtue of their being encoded with terminologies and ontologies, which could be leveraged to assist with this process and provide a greater level of certainty.

3.3.2 *Method*

For pathway analysis to be effective and meaningful, a raw health record needs to be filtered down to a manageable number of events that are relevant to the particular research question. This section outlines a method for automatically generating and abstracting codelists through the use of ontologies, and a framework for evaluating these codelists. Given a particular research scenario, a researcher should be able to specify a basic concept, for example *colorectal cancer*, and receive a list of relevant procedure codes with which to filter data ready for process analysis.

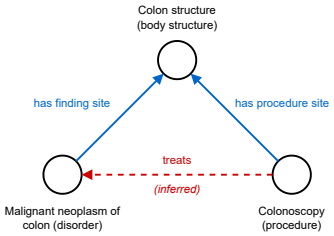

Figure 3.10. Inferring a new *procedure treats disorder* relationship from two known relationships

Query 1. "Which concepts are procedures, and occur on a site that is a possible finding site for our target diagnosis?"

Identifying pathway events using ontologies

A structured query for identifying concepts of interest relevant to a particular disease area was defined in SNOMED's Expression Constraint Language (ECL). The basic intuition behind this method is simple: given a procedure P, an anatomical site S, and a disease D, if the relationships (P, *procedure site*, S) and (D, *finding site*, S) exist in the ontology, it can be inferred that P is potentially a treatment for D (Figure 3.10). It is not guaranteed to be an acceptable or recommended treatment, but this logical link means that it is at least plausible. The idea is that this simple rule should be able to eliminate a large number of the procedure codes in a dataset from consideration.

This logic can be formulated in ECL as the following query:

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

In this notation, the < operator retrieves all concepts that are a descendant of another, << retrieves all descendants plus the original concept itself, * is a wildcard, indicating any concept, : indicates refinement, and R indicates a reversal - i.e. a retrieval of the set of attribute values that exist given a particular attribute and a set of concepts. The term `71388002|Procedure|` identifies the SNOMED concept *procedure*, which has the ID `71388002`. Thus, this command retrieves, from the set of all descendants of *procedure*, concepts whose attribute *procedure site* is equal to, or descended from, any concept within the set of possible *finding sites* for the `Diagnosis` SNOMED concept provided by the user.

The search starts with a specified diagnosis, rather than another concept type such as the site, because this is the level at which research questions are typically focused (e.g. "what do treatment pathways for colorectal cancer look like?"). A wide range of further constraints are possible: based on the results of this basic query, the exact formulation is improved and iterated upon.

Transforming pathway events using ontologies

Once a set of relevant SNOMED concepts has been retrieved, they need to be converted into the terminology used by retrospective datasets, and into a readable and useful form.

The first step is mapping between code systems: convert the output of the SNOMED ECL queries to the ICD and OPCS form used by the data. In initial studies (Dwyer et al. 2024), the National Health Service's official SNOMED to

OPCS mapping files were used, but these came with some downsides. These maps are intended to assist clinical coders in converting codes for billing purposes. Most SNOMED concepts map to either a choice of multiple OPCS codes, or combinations of codes, with the expectation that clinical coders make the choice. This is acceptable for filtering purposes, since the main aim is to establish plausibility ("could this event be related to this disease?") rather than to guarantee causality ("this event must have been intented as a treatment for this disease"). Given the precedent that questionable codes should be retained unless there is a clear rationale for exclusion (Watson et al. 2017), the assumption was made that any OPCS code that is mapped in any way to a SNOMED concept could plausibly be related. However, this method led to relatively low query precision, and meant that extra maps were required to translate to ICD-9 Vol. 3 and ICD-10-PCS codes, where some obvious gaps and missing links were identified.

To address these issues, the experiments described here used a different approach: SNOMED concepts were mapped to their OPCS and ICD equivalents using the OMOP Common Data Model (Reich et al. 2024). Previous research has shown that UK format EHR data can be converted to this data model, with 99% of OPCS concepts being covered (Papez et al. 2021, 2023).

The second transformation step is code abstraction. A major issue with healthcare PM is the complexity of the process models created, in particular the propensity for so-called "spaghetti models". There are over 10,000 possible events in the OPCS code system; even after filtering to just the relevant ones, we can still be left with a very large number, creating a need to group them together into meaningful and interpretable categories. Whilst there are several ways to achieve this, this experiment used the code's text descriptions to identify natural groups. A set of keywords was established based on the groups used in the colon and lung cancer lists, and these were used to group codes into categories. For example, any concept that contained 'resection', 'excision', 'lesion' or 'exentoration' in its description was mapped to 'resection'.

3.3.3 *Evaluation*

Three approaches are used to evaluate the quality of results from the ECL queries: a comparison of the output codelist against established research codelists, a comparison of the resulting process models against models created from benchmark codelists, and a comparison of the codelists against the statistical distribution of codes in the data.

Evaluation against codelists

Firstly, each output codelist is compared against established codelists from medical research studies. To ensure that we gain a full picture of our methods' effectiveness, and how that effectiveness varies by disease, codelists were taken from three different sources, and were of varying sizes (Table 3.1).

Codelists A, B, and C, which represent appendicits, cataract, and glaucoma, come from the HDRUK Phenotype Library. These were chosen from the relatively small number of available codelists which contain procedural codes, and priority was given to longer codelists in order to produce more meaningful precision and recall statistics (for example, diverticular disease provides only one OPCS code, which would make evaluation meaningless). These three codelists all originate from the same publication (Kuan et al. 2019), in which the presence of particular OPCS codes (in addition to or in place of other diagnostic and procedural codes) are used to determine whether a patient has or does not have a particular condition. Codelist D was constructed based on lung cancer procedures from consultation with domain experts. Codelist E comes from an established research database, the COloRECTal cancer Repository (CORECT-R; Downing et al. 2021). This project has published an extensive codelist of 192 different OPCS codes, divided into five treatment categories, that cover the main treatments for any colon or rectal cancer (*CORECT-R Data Coding* 2020). A previously prepared extension of this codelist was used, which adds several additional codes for relevant diagnostic tests, chemotherapy, radiotherapy, and surgical procedures to gain a full picture of the entire patient journey for colorectal cancer.

List	Disease	*n*
A	Appendicitis	14
B	Cataract	32
C	Glaucoma	21
D	Lung cancer	76
E	Colorectal cancer	497

Table 3.1. List of benchmark codelists used and their length (*n*)

These codelists come from different sources and serve different purposes: A, B, and C act as minimum thresholds regarding whether or not a patient should be considered to have a particular disease, 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 this second set of codelists are designed to requirements more closely resembling those needed for PM; this range of different sources is deliberately included to consider how well this approach generalises to different scenarios.

The lists are used to evaluate the output of each ECL query according to precision (the proportion of retrieved codes that were in the reference list) and recall (the proportion of codes in the reference list successfully retrieved).

Comparing processes

To investigate the extent to which our methods change the results of analyses, codelists produced by these queries are used to generate directly-follows graphs of treatment processes from real-world datasets. The dataset analysed was collected by Oxford University Hospitals (OUH) NHS Foundation Trust, which operates four hospitals in Oxfordshire, England. From the raw health records of patients diagnosed with either lung (ICD-10 codes C33*, C34*) or colon (C18*) cancer, two directly-follows graphs of events for each disease were produced: one filtered from a benchmark OPCS codelist, and one filtered according to an OPCS codelist derived from our SNOMED ECL queries.

Comparing statistics

A final useful method of evaluating the proposed approach is to compare it against a statistical approach, in which the relevance of events is determined based on their correlation with a particular diagnosis. The set of correlated events can be compared to the events from logically constructed codelist, making it possible to identify codes that have no clear logical link to the target disease but are still associated with a diagnosis according to the data.

However, a limitation of the OUH dataset is that the data available only consists of patients with colorectal or lung cancer, making it impossible to see which codes are over-represented in the colon and lung cohorts compared to the general population. Therefore, I also make use of MIMIC-IV (Johnson et al. 2023a,b), a freely accessible EHR dataset from Massachusetts, USA which has been widely used in previous PM studies. In MIMIC, each procedure is characterised by an ICD code, using a mixture of ICD versions 9 and 10, which can be converted to from SNOMED using the same OMOP maps used to map OPCS codes (Section 3.3.2).

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, we identify codes with a statistically significant difference in frequency between the two cohorts, and therefore some association with a diagnosis of colon or lung cancer.

3.3.4 *Results*

This section describes the results of the proposed method by comparing the quality of the generated codelists against reference lists, by comparing the process models generated using these lists, and by examining the distribution of codes in a supplementary dataset.

Query development

Initially, concepts were retrieved based on the query outlined in Section 3.3.2 (referred to here as Query 1). During initial tests, this query retrieved many codes which were only tangentially relevant: applying this example to colon cancer, for example, returned a number of codes describing *oesophagectomy*, the surgical removal of the oesophagus. Because this procedure sometimes involves replacing the oesphagus with material from the colon, several concepts therefore include the colon as a procedure site. In practice, this procedure is very unlikely to be related to colon cancer; the oesophagus is the primary site of this procedure, and the colon is of secondary concern. Therefore, to exclude such concepts, an additional cardinality constraint was introduced, ensuring that procedures must have one and only one site:

Query 2. "Which concepts are procedures, occur on a site that is a possible finding site for our target diagnosis, and have exactly one procedure site?"

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

A second issue with Query 1 was that it also retrieved some procedures that were seemingly totally irrelevant. For example, in some older releases of SNOMED CT, the concept representing *malignant neoplasm of colon* had child concepts including *malignant neoplasm of rectosigmoid junction metastatic to brain*, which meant that the brain was also automatically introduced as a procedure site of interest, so any number of codes relating to the brain are returned as a result. In case such examples exist within other diseases, a more specific query was created: instead of requesting any possible site for any child concept of the target disease, it requests only procedures that involve a specific, named anatomical concept:

Query 3. "Which concepts are procedures, and occur on a specified site?"

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

Finally, the new rules introduced in Queries 2 and 3 were combined to constrain our output to those procedures with exactly one procedure site, *and* a procedure site that is a specific named site, or a descendant thereof:

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

Query 4. "Which concepts are procedures, occur on a specified site, and have one and exactly one procedure site?"

When running these queries on cancer diagnoses, there also exist several other procedures that are highly relevant but not site specific: for this reason, an additional request for chemotherapy and radiotherapy codes was appended onto relevant queries:

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

Query 5. Additional rules to capture chemotherapy and radiotherapy concepts

Comparing codelists

Figure 3.11 compares the results of Queries 1–4 in the five different disease scenarios. The queries were run against the February 2024 release (`37.5.0_20240214000001`) of the SNOMED CT UK Edition. OPCS codes in chapters Y and Z — supplementary codes that describe additional information such as sites or approaches to surgery — are excluded from calculations since as a rule they never appear in the primary position in records, and therefore have little bearing on process mining results.

In terms of raw codelist size, moving from the more general Query 1 to the more restrictive Query 4 consistently reduced the number of codes returned, as expected. However, two diagnoses — cataract and glaucoma — produced exceptionally large result sets for Queries 1 and 2. This occured due to the existence of concepts in SNOMED that represent multiple diagnoses, for example the concept *glaucoma and sleep apnea* means that *structure of the eye* and *structure of the respiratory system* are both introduced as acceptable procedure sites, so any procedures on the respiratory system are introduced, significantly inflating the codelist. This was significantly improved by the introduction of the more narrow site definition in Queries 3 and 4.

Precision was generally low. This was again most visible in the cataract and glaucoma results, where the exceptionally large codelist size made it inevitable that a very small number of those codes would be relevant. However, it is also noticeable that in the case of the cataract codelist, and to a lesser extent in appendicitis, Queries 3 and 4 represented a significant jump in precision, suggesting that the move from a specific disease to a single named procedure site did successfully eliminate a large number of irrelevant

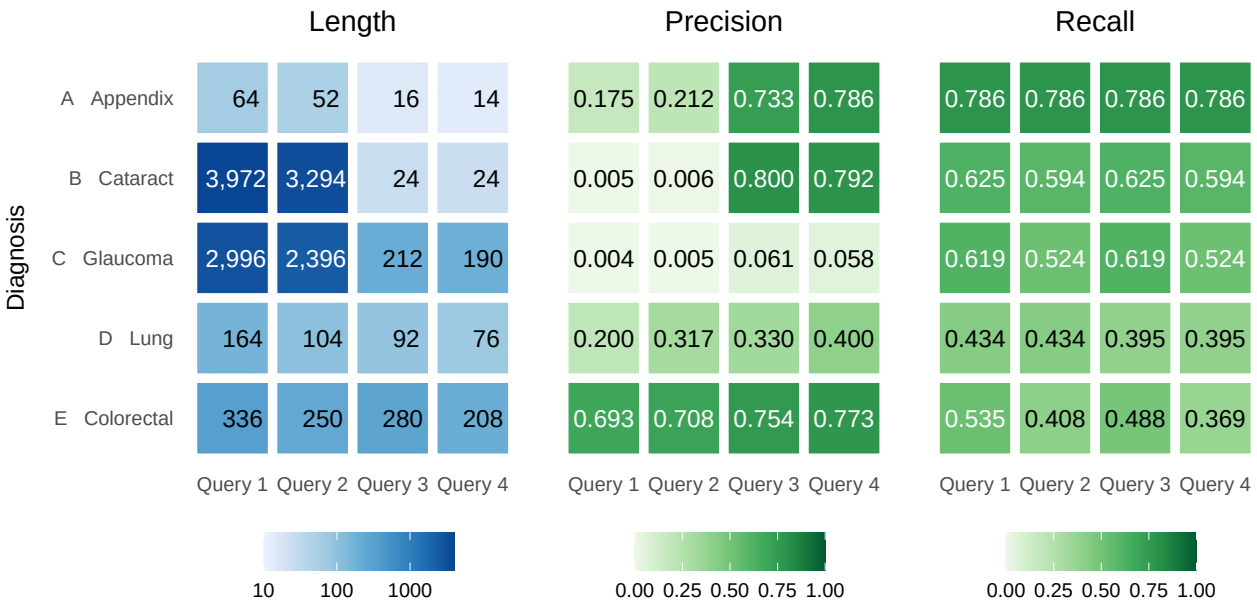

Figure 3.11. Length, precision and recall of codelists resulting from Queries 1–4 against five different target diagnoses

concepts. Colon cancer was the notable outlier in terms of precision, likely owing to the significantly larger size of the benchmark codelist compared to the others.

Recall, by contrast, was relatively high across all scenarios tested. For appendicitis, it was entirely constant across all queries, indicating that the changes to constraints did not remove any correct codes. In every other case, recall was lower in Query 4 than in 1, indicating that increasing constraints — while effective at reducing irrelevant codes — also inevitably resulted in a loss of some relevant codes.

Figure 3.12 shows the proportion of concepts in the benchmark codelist retrieved, with retrieved codes grouped into categories. Concepts representing blood tests, diagnostic imaging and testing and patient assessment were never retrieved, because they are not disease specific and therefore not annotated with a clear site in the ontology. Similarly, endoscopy of the upper gastrointestinal tract was never retrieved as it does not have a relevant procedure site, but it is commonly included in colon cancer analyses as it commonly forms part of the route to diagnosis. Recall was highest in surgical topics, since they were well annotated with clear procedure sites in the ontology, and the increased constraints moving from Query 1 to 4 only had a noticeable affect on these codes.

Comparing processes

To investigate the usefulness of these codelists for PM, directly-follows graphs for colon and lung cancer patients were generated from the OUH

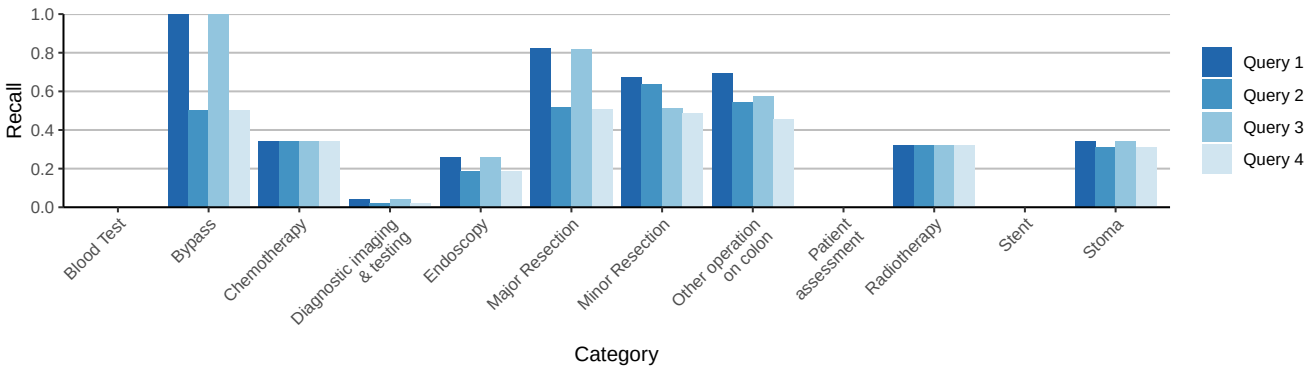

Figure 3.12. Recall of Queries 1–4 on colon cancer against the extended CORECT-R reference set, grouped by category

dataset, based on the generated codelist according to Query 1 (chosen for consistently displaying the highest recall) and the known benchmark codelist. In these graphs (Figures 3.13 and 3.14), node labels represent relative case frequency, i.e. the number of patients experiencing that event, whilst edge labels represent the relative antecedent frequency, i.e. the proportion of instances of the source event that are followed by the destination event.

In colon cancer, the *minor* and *major resection* were grouped into one *resection* category for clarity and simplicity, as were *lower gastrointestinal* and *upper gastrointestinal endoscopy*. Activities that only occured in a very small proportion of patients (stoma, stent, bypass, other, imaging, assessment) were excluded.

The two processes resemble each other in many ways. A similar proportion of patients underwent resections (28.23% in our codelist, 36.9% according to the benchmark codelist), reflecting the a reasonably high recall for resection-related codes. Both processes share similarities: the most common first event was resection; imaging and radiotherapy were typically the last event in a pathway; and radiotherapy very rarely follows a resection. However, there are also some significant differences: the proportion of patients experiencing chemotherapy, imaging and endoscopies were all markedly lower in our model than the CORECT model. Additionally, 59.61% of patients did not experience any events of interest, compared to 33.12% according to the benchmark list.

In lung cancer, there was a higher level of agreement between the two process models. In particular, a very similar rate of lobectomies were recorded by both models, indicating that both methods arrived upon similar codelists in this category. Chemotherapy and radiotherapy rates were also similar. The smaller scale of discrepencies between these two models reflects the fact that the number of possible codes for lung cancer are considerably fewer than for colon cancer, given that treatment via surgery is less common.To investigate the usefulness of these codelists for process mining, directly-follows graphs

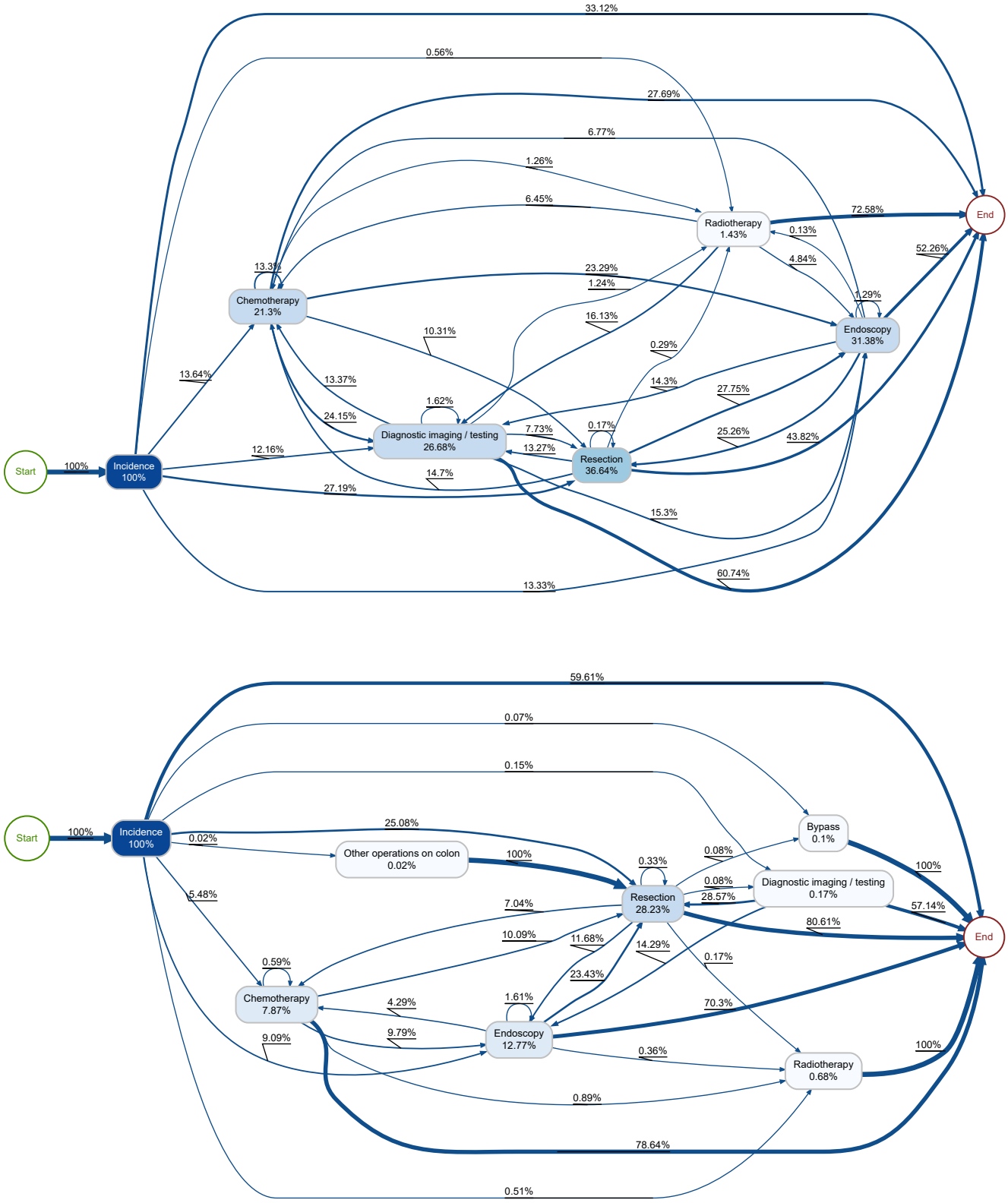

Figure 3.13. Directly-follows graphs summarising the pathways of colon cancer patients, filtered according to the benchmark codelist E (top) and Query 1 (bottom)

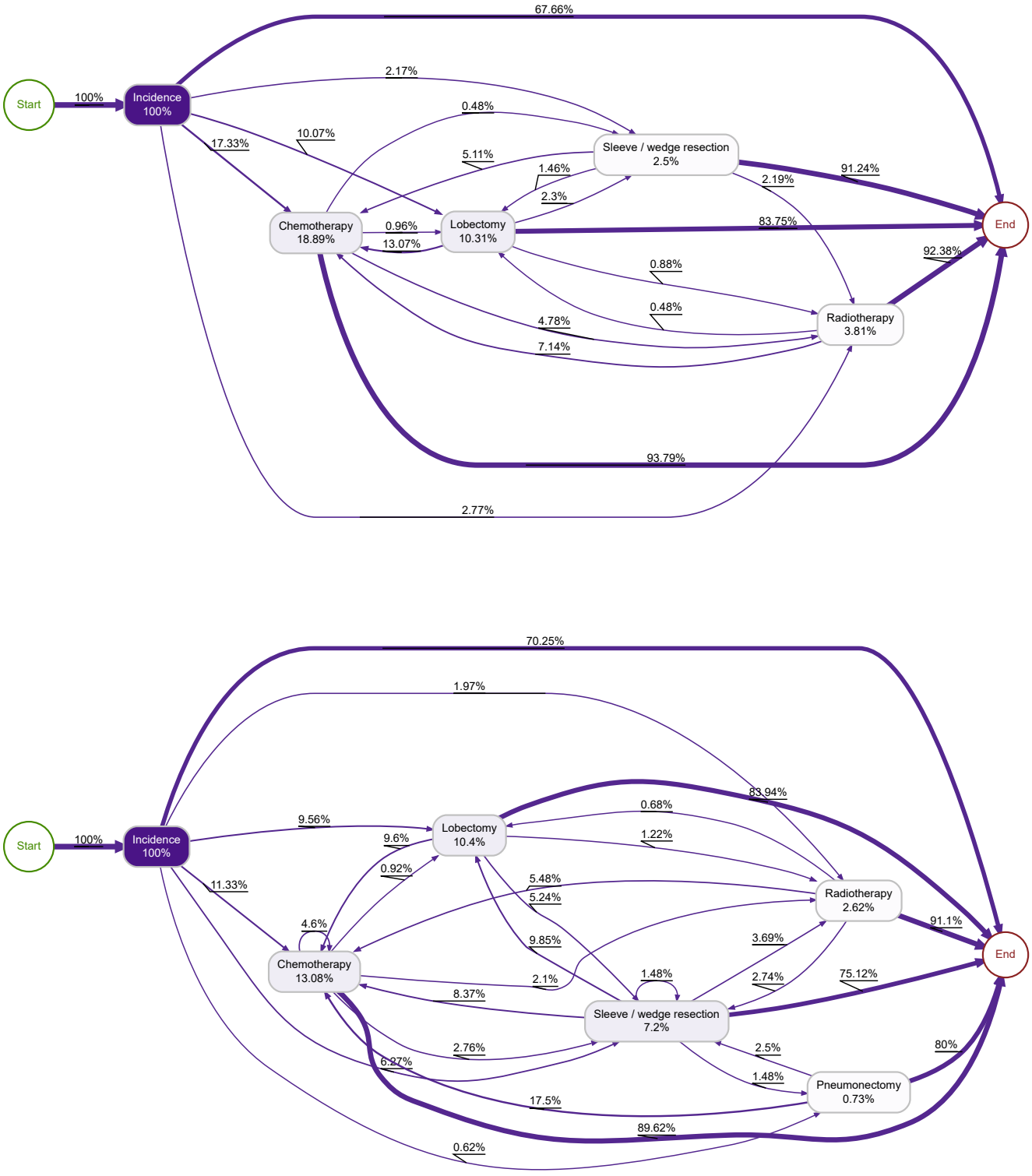

Figure 3.14. Directly-follows graphs summarising the pathways of lung cancer patients, filtered according to the benchmark codelist D (top) and Query 1 (bottom)

for colon and lung cancer patients were generated from the OUH dataset, based on the generated codelist according to Query 1 (i.e. the highest recall) and the known benchmark codelist.

Comparing with statistics

The procedure codes with the strongest association with a colon or lung cancer diagnosis within the MIMIC dataset are shown in Table 3.2. Each code was counted in the diagnosed and non-diagnosed population, and the rate of occurance compared by calculating an odds ratio and statistical testing. Listed in this table are those codes that were more common in the case cohort, appeared > 10 times in the case cohort, and were statistically significant ($p < 0.05$) according to Fisher's exact test.

For colon cancer, most of the missing codes related to the large intestine, and therefore were not found by our queries as they were not tagged with a descendant of "colon" as a procedure site ("colon" is itself a descendant of "large intestine" in SNOMED). The exception here is *endoscopic insertion of colonic stent(s)*, which does not have a recorded map to a particular SNOMED code in the OMOP data model, although it does have an *is a* relation to a similar SNOMED concept. Lung cancer follows a similar pattern, with many of the codes being related to the lymphatic system and therefore not having a relevant procedure site tagged. *Endoscopic excision...* again did not have a map to a directly equivalent SNOMED concept.

3.3.5 *Discussion*

These results suggest that the proposed approach to preparing event logs using queries over ontologies can approximate an expert-prepared codelist. There is a clear trade-off between precision and recall: depending on the query structure, the results lie on a spectrum between a relatively complete codelist with a large number of extra codes, or an incomplete codelist in which every code is likely be relevant.

Performance varies depending on the type of event: surgeries are often easy to identify, owing to their clear structure and the prevalence of useful annotations in the SNOMED ontology, whereas relevant but non-disease specific events such as chemotherapy, diagnostic tests and imaging need to be explicitly specified. In some cases, there exist relevant concepts that don't have a clear ontological relationship to known concepts: procedures in adjacent locations can happen due to surgical complications or cancer metastasis. However, any widening of definitions needs to be balanced with

ICD concept			Prevalence		Odds ratio	p
Version	Code	Description	Control	Case		
		Colon				
10	0DTF4ZZ	Resection of Right Large Intestine, Percutaneous Endoscopic Approach	0.05	6.69	137.57	7.88e−108
10	0DTG4ZZ	Resection of Left Large Intestine, Percutaneous Endoscopic Approach	0.01	1.15	82.53	6.76e−18
9	4593	Other small-to-large intestinal anastomosis	0.15	7.65	55.54	1.39e−99
10	0DTF0ZZ	Resection of Right Large Intestine, Open Approach	0.07	3.25	51.38	1.41e−42
9	4594	Large-to-large intestinal anastomosis	0.08	2.87	34.97	1.27e−33
9	4686	Endoscopic insertion of colonic stent(s)	0.05	1.53	31.29	5.14e−18
10	0DJD4ZZ	Inspection of Lower Intestinal Tract, Percutaneous Endoscopic Approach	0.06	1.15	20.93	4.51e−12
9	4579	Other and unspecified partial excision of large intestine	0.06	1.05	16.46	3.49e−10
10	0D1B0Z4	Bypass Ileum to Cutaneous, Open Approach	0.18	2.58	15.08	6.71e−22
9	4620	Ileostomy, not otherwise specified	0.12	1.43	12.46	8.89e−12
		Lung				
10	07T70ZZ	Resection of Thorax Lymphatic, Open Approach	0.00	0.90	361.96	5.75e−43
10	07B70ZZ	Excision of Thorax Lymphatic, Open Approach	0.01	1.12	122.81	1.32e−47
10	07B70ZX	Excision of Thorax Lymphatic, Open Approach, Diagnostic	0.01	1.53	115.38	1.02e−63
10	07B74ZX	Excision of Thorax Lymphatic, Percutaneous Endoscopic Approach, Diagnostic	0.07	4.74	66.48	6.81e−176
9	3228	Endoscopic excision or destruction of lesion or tissue of lung	0.03	1.90	65.91	1.60e−71
9	3422	Mediastinoscopy	0.05	2.68	59.59	5.34e−98
9	3230	Thoracoscopic segmental resection of lung	0.02	1.12	56.28	1.96e−41
10	07D78ZX	Extraction of Thorax Lymphatic, Via Natural or Artificial Opening Endoscopic, Diagnostic	0.03	1.25	50.09	1.17e−44
10	07T74ZZ	Resection of Thorax Lymphatic, Percutaneous Endoscopic Approach	0.02	1.15	49.59	2.50e−41
10	07B74ZZ	Excision of Thorax Lymphatic, Percutaneous Endoscopic Approach	0.04	1.56	44.86	6.08e−54

Table 3.2. ICD9Proc and ICD10-PCS procedure codes that were strongly associated with a colon or lung cancer diagnosis (p<0.05, ranked by odds ratio) and did not appear in the Query 1 generated codelist

the risk of increasing the number of irrelevant codes. Query constraints always need to be guided by the research question and purpose.

Performance also depends on the choice of, and the original purpose of, the benchmark codelist. Codelists A, B, and C were defined by their original authors for the purpose of selecting a patient cohort, rather than to explicitly study the care processes of those patients, which is the purpose of Codelists D and E. Codelists are designed with specific intents in mind, and it is important for process mining researchers to bear this in mind when reusing them for comparison.

Our results also highlighted that abstracting events into meaningful and useful categories is still a key challenge. When developing our methodology, three approaches were considered. One option was using the structure of the data's original terminology. OPCS codes consist of four characters, which describes three hierarchical levels. 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. Aggregating codes using this structure was in practice challenging. Since OPCS chapters are arranged at the highest level by site (with the exception of the *diagnostic imaging, testing, and rehabilitation* and *miscellaneous* chapters), almost every code related to a particular disease will come from one chapter, resulting in almost all events being aggregated together. 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 for event aggregation. In the example of Query 1 for colon cancer, the results span 80 three-character concepts, each containing at most 10 codes, which is far too many to create an interpretable process model. Therefore, the level of abstraction means that the OPCS terminology structure in and of itself is not a useful framework for event aggregation.

A second approach is to aggregate the codes before their transformation, within the SNOMED 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 methods *repair* or *excision* as surgery. By examining the most common *methods* relationships that exist within the codelist, a set of maps that take advantage of this structured knowledge were hand-prepared. In this case, the main barrier to useful aggregation was the fact that many OPCS concepts were mapped to multiple possible categories in the SNOMED terminology. In most cases this was because the source SNOMED concept had multiple *method* relationships, but occasionally because one OPCS concept could be mapped to multiple SNOMED concepts. Additionally, the high number of possible *methods* created a very large number of categories; in the case of Query 1 for colon cancer, there were 48.

A major issue shared by both the OPCS and SNOMED-level aggregation approaches is that the categories do not clearly align with those from the benchmark codelists. This means that any resulting process models are not comparable with those from a different source codelist. For this reason, I used the text descriptions of each code, defining a set of categories equivalent to those in the colon and lung cancer benchmark lists, and matching codes to these based on a keyword search: These keywords were chosen based on patterns identified from manually examining the benchmark codelist, to ensure as close an alignment as possible. 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 categorisations from comparable codelists.

It is therefore clear that at present, the structure of the key ontologies used in healthcare data do not easily support aggregation into a useful level, and some level of human curation is realistically still required. Effective abstraction for research also depends on the research question: modelling an overarching treatment process requires a different abstraction level to trying to determine smaller process differences within a particular treatment pathway, for example.

The proposed approach is capable of providing a reasonably accurate list on which to begin domain expert-lead discussions, speeding up the initial stages of the development process. It allows the majority of concepts in a codelist to be described through as a set of intuitive and explainable inclusion criteria in a standard language. As a result, the codelists themselves consist of concepts from a standard ontology that can then be translated into the languages used by individual datasets — although this is highly dependent on both the completeness of the original ontology and the quality of mappings

available. This approach could also be informative in situations where very small amounts of data are available, limiting the usefulness of statistical analysis of the most common codes. Finally, the process of specifying a set of requirements for desired events in terms of their relationship to a particular target disease keeps the PM method focused on a concrete health issue, ensuring clinical meaningfulness and good alignment between PM studies and current research questions in healthcare. Domain knowledge, even in the form of well-established and curated ontologies, is rarely perfect and cannot instantly automate the process of event log preparation. However, it can provide a meaningful framework to generate ideas to be iterated upon, and inspire conversations with domain experts.

3.4 *Summary*

This chapter presents two views on patient pathways. Firstly, from a healthcare point of view, I consider what a clinical pathway actually is, and what information is needed to properly interpret them in context. Secondly, I examined pathways from a practical and computational perspective, asking what clinical pathways look like in real data, and how the right data should be extracted to analyse them in a systematic way.

The first major contribution of this chapter is a demonstration of the value of context in interpreting pathways. Through examining the real journeys of patients with colorectal and lung cancer, I found that a simple view of pathways is rarely enough to draw meaningful conclusions. However, considering contextual factors provides significantly more interesting results that align with expert knowledge of pathways, and with results previously healthcare literature. The second contribution is a proposed approach to automatically extracting and preparing healthcare data for pathway analysis, which makes use of domain knowledge from ontologies. The results found that this approach to event extraction is possible, and that codelists generated through this method resemble those that would be prepared by human experts; that event aggregation is more challenging, due to the relatively rigid options provided; and that taken together, when used to produce process models, the resulting process models do resemble those prepared with more traditional methods.

The development and design of these studies also revealed several interesting points around the structure of health data. For example, Section 3.2 defines a patient of interest as a patient who has a recorded diagnosis of the target disease within the cancer patient management system, whilst Section 3.3, by contrast, includes any patient who has the target diagnosis recorded

at some point, resulting in a much larger cohort. The former definition is almost certainly the preferable one: early analyses included a very high proportion of patients with no treatment at all — a pattern visible in Section 3.3 — which is likely due to the recording of clinical codes to represent suspicion or investigation rather than diagnosis. For this reason, the more specific definition was devised based on knowledge of the hospital's information system, specifically the fact that a separate system is used specifically for tracking cancer patients. This was not possible to incorporate into Section 3.3 as this batch of data was not available at the time, or for lung cancer patients, but the difference in these two sections' findings, especially regarding patients supposedly going without treatment, underlines a key point of this chapter: that EHR data is heavily influenced by human practices and is not a perfect record, and that a strong understanding of these processes is needed to perform meaningful analysis.

Collectively, these findings emphasise the benefits of structured ontological knowledge and statistics in enhancing and expediting research, but fundamentally support approaches that are guided by domain expertise. The automated approach is far from a replacement for discussion and input from domain experts; it in fact underlines the importance of a deep understanding of the subject area, of the research question, and of the very nature of electronic health record data. They underline the importance of a patient-centred and non-prescriptive approach to healthcare process modelling. Variations in care are not "noise" or "anomalies" to be ironed out, but part of an individual patient's story, and decisions that appear unusual on paper may have been made according to clinically sound judgement.

Thus far, attempts to apply process mining methods to healthcare data have not always been successful. This is because these methods are designed for business processes. This chapter argues that patient pathways are not deterministic production lines, but by necessity unique to each patient. Analysis such as deriving average pathways, or grouping pathways into clusters are useful, in that they allow us to explore population-wide patterns and make sense of very large populations, but they still need to account for this natural heterogeneity.

4 *A distance measure for patient pathways*

Given the complexity of pathways, dividing patients into binary groups based on "compliance" and "non-compliance" is overly simplistic, and conceals heterogeneity within groups. In this chapter, I therefore outline a method for assigning continuous values that describe the difference between patient histories. This proposed method has several notable characteristics, which are deliberately chosen to avoid the pitfalls of many modern machine learning approaches, and to emphasise model practicality, accessibility, and reproducibility.

By combining knowledge graph embeddings with timeseries methods, it allows the description of distance between patient histories based on both the semantic similarity of the events and their temporal patterns. By using embeddings trained only from an internationally standardised and publicly available domain ontology, the risk of accidental leakage of patient records is removed, and embeddings are created that can be publicly shared, allowing for the reproduction and comparison of results. By training the core embedding model on the entire SNOMED ontology, rather than a particular medical area or a particular dataset, embeddings are created that only need training once, and can be reused for different research questions; this significantly lowers the computational power and cost requirements, and reduces the environmental impact of performing analyses.

I compare several existing KGE algorithms, and use them to generate embeddings of the SNOMED ontology (Section 4.2), specifically investigating whether performance on traditional link prediction metrics necessarily entails an effective representation of semantic similarity.

I then introduce embedding-based dynamic time warping (E-DTW), a method for combining concept embeddings into a single representation of a patient's pathway and measuring the distance between pathways (Section 4.3). The proposed method represents a patient's pathway as a timeseries of points in an embedding space and measures the distance between pathways with a dynamic time warping-based algorithm. This method is evaluated on its ability to describe differences in patient traces, and is applied to a retrospective EHR dataset to examine its effectiveness in describing patient differences in specific research scenarios (Section 4.4).

4.1 *Background*

Ontologies encode curated and logically structured domain knowledge, which makes them exceptionally useful for a wide range of applications. However, like all non-numeric data, this knowledge is difficult to actually analyse with modern machine learning methods which typically require numerical input. For this reason, a wide range of methods have been developed that aim to convert graphical knowledge into low-dimensional representations, whilst preserving as much of the original structure as possible. These *knowledge graph embedding* (KGE) approaches are also known for typically placing similar concepts close to each other in this low-dimensional space, although this is not necessarily guaranteed.

4.1.1 *Embedding methods*

Many of the foundational KGE approaches have their roots in *word embedding*, the task of converting free-text data into embeddings. One of the seminal algorithms in word embedding is word2vec (Mikolov et al. 2013b). This relatively compact (two-layer) neural network introduced the *skip-gram* model, which emphasises finding a representation of a word that can predict the contents of a sliding window of surrounding words, giving higher weighting to closer words. Word2vec's simplicity and accessibility means that it has been highly influential, and has been widely modified and applied to non-textual data. Kmer2vec (Ren et al. 2022), for example, learns embeddings of long DNA sequences by splitting them up into k-mers — substrings of a specified length k — and feeding them into word2vec as "words". Node2vec (Grover & Leskovec 2016) and RDF2vec (Ristoski & Paulheim 2016) work similarly, learning representations of both simple graph structures and knowledge graphs with semantic edges by treating random walks along the graph structure as sequences of words. Most relevantly for our purposes however, word2vec's principles have also been widely applied to learning embeddings of medical codes (Beam et al. 2020; Choi et al. 2016).

Since then, more complex word embedding methods have been proposed, notably GloVe, which emphasises global rather than local co-occurrence; and ELMo and BERT, which use deeper neural networks to give the same word different embeddings depending on context. Getzen et al. (2024) compare all of these models, and find that they all display roughly equivalent performance on prediction tasks that require structured medical concepts, except for ELMo and BERT which are notably inferior. This is perhaps surprising, as these are more recent and more complex models, but the authors note that they

require a substantial volume of training data that was not available, which was likely a factor in their performance.

What all these approaches have in common is their reliance on the *distributional hypothesis*: the idea that a word's meaning is characterised by those words it frequently co-occurs with, an idea which has a long history in philosophy and linguistics (Skelac & Jandrić 2020). The application of this principle to increasingly complex models and increasingly large training datasets has culminated in modern LLMs. The work of Kane et al. (2023) is perhaps the natural evolution in text-based medical concept embeddings, describing the use of an LLM to create embeddings of ICD-10 concepts. However, this work highlights several disadvantages which extend to any language-based approach: the reliance on quality training data, in particular the challenge of gathering enough input data to effectively handle less common concepts; the difficulty of representing meaningful hierarchies such as those found in ontologies; and the risk of overfitting to the training dataset.

Overall, language models of various scales have been widely used to learn representations of words and other concepts, but their shortcomings are particularly limiting in healthcare contexts. They rely on measuring co-occurrence and statistics which — whilst undoubtedly effective — make them less useful the rarer a particular concept is, and they cannot guarantee representation of semantics or structure. These models can be adapted to work on graph data, however this typically only considers the simple graph structure, and not relation semantics, treating relationships as simple indicators of connectivity, rather than carriers of meaning. The distributional hypothesis is well-founded: medical concepts' meanings are undoubtedly influenced by the contexts in which they appear, and language models capture this well, but it is also true that medical data and healthcare processes are very heavily influenced by other contexts which are not always visible in the data. Specific patient populations, healthcare systems, and human processes all impact what data is recorded, how it is is recorded, and the patterns that it exhibits. This means that a model trained on a different population will not be able to guarantee generalisability. Finally, the reliance on real healthcare data as a training source also means that it is exceptionally difficult to share this training data publicly, which severely limits the reproducibility of these models.

In parallel, a separate class of models known as knowledge graph embeddings (KGEs) have developed, specifically designed to create embeddings of knowledge graphs. By explicitly developing a mapping between entities and relations and their vector counterparts, they aim to more accurately describe the characteristics and relationships between them (Biswas et al.

2023). These methods are introduced in depth in Section 2.5. In practical terms, a KGE algorithm typically works in three steps. Firstly, entities are represented as points in a vector space, and relations as some kind of operation within the space, which can be described with a vector, matrix, tensor, or probability distribution. Secondly, the representations are then evaluated in terms of plausibility using some scoring function, which scores facts that actually exist within the KG higher than those that do not. Finally, these representations are iteratively improved; this is an optimisation problem that aims to maximise the plausibility score (Wang et al. 2017).

Knowledge graph embeddings rely on structured, curated knowledge rather than text corpora, and therefore present an alternative, and arguably under-studied, approach for learning representations of medical concepts. The body of work on KGE-based representations of medical concepts is slimmer than that for text models, but Chang et al. (2020) evaluated five different translational KGE models alongside two skip-gram based models on entity classification and relation prediction tasks and found that the KGE models almost always outperform the word2vec-inspired ones, indicating that these methods hold promise.

Ontologies are not necessarily a perfect data source: they are curated by humans and so can contain logical contradictions and inconsistencies (Mortensen et al. 2015; Slater et al. 2020), or encode outdated or discriminatory concepts (Ram et al. 2022). The difference is that in an ontology, an incorrect fact or chain reasoning can be traced to its source and corrected.

4.1.2 *Evaluating embeddings*

Knowledge graph embeddings are typically evaluated on the task of *link prediction*, i.e. how well the embedding is able to predict the presence of unseen relationships between entities, withheld during the training process. However, whilst link prediction is by far the dominant evaluation metric, it is not the only one. Portisch et al. (2022) identify two main tendencies in KGE research: some embedding approaches prioritise link prediction, the primary aim being to distinguish between correct and incorrect facts using the graph as a source of ground truth, whilst others prioritise data mining, aiming to produce low-dimensional representations of graphs as input for other machine learning algorithms.

These are often considered to be two very different tasks. Link prediction approaches are often said to encode semantic similarity into the embedding space, with entities in the same category forming natural clusters, but this is a side effect of optimising for a particular prediction operation rather than by

design. If two entities are considered to be similar when they share a common relation to a common third entity, then it follows that this will create natural clusters of similarity. Data mining approaches, by contrast, will position embeddings close to each other when they share *any* relation to a common third entity, or a common relation to two different entities, meaning that whilst these models aim to encode *similarity*, they also encode *relatedness*. Portisch et al. find that link prediction methods can perform reasonably well on classification and regression tasks, and data mining models can perform well in link prediction, indicating that the line between these two paradigms is less clear-cut than one might expect.

However, other sources have questioned the semantic meaningfulness of knowledge graph embeddings. Jain et al. (2021) compare the performance of KGE representations with alternatives generated via SDType, an approach which calculates the probability of each entity belonging to a particular type based on the types of its incoming and outgoing relations (Paulheim & Bizer 2013). Whilst this is used to argue that a "traditional" method can outperform KGEs, the results show that the alternative approach is in most cases roughly comparable to KGEs, is better on one or two occasions, and is the worst approach on another one or two occasions. Additionally, this method is highly specific to classification, and is only evaluated based on classification performance, rather than on clustering or any other metric of semantic similarity. Therefore this proposed alternative, whilst capable of performing comparably to KGEs, only does so for the specified problem for which it is designed; by contrast KGE representations are reusable as inputs to a many algorithms once generated. The key takeaway from these experiments is the sheer range of results, which highlight that there is likely no "best" KGE model, and different embedding models can perform very differently to each other on different datasets. Other research has indicated that changes in training strategies and parameters can also have a substantial effect on model quality (Gema et al. 2024; Ruffinelli et al. 2019).

Hubert et al. (2024) similarly question the effectiveness of KGE methods, arguing that since KGE models are typically trained to optimise performance in link prediction scenarios, the proximity of entities in an embedding space is not guaranteed to align with semantic similarity. As previously noted, performance in classic rank-based metrics is not totally separate from entity similarity: the principles of distributional semantics suggest that similar entities will naturally appear in similar triples as a side effect, meaning that there is likely some correlation between rank-based metrics and entity similarity, but this work attempts to quantify the level to which this occurs. The proposed method compares a concept's set of immediate neighbours

in the original graph with that concept's set of closest embeddings, and measures the degree of overlap. By this metric, the TuckER and DistMult KGE models display the best similarity, and BoxE, RDF2Vec, RESCAL and TransE the worst, but as before, performance varies substantially between datasets and even between different classes within the same dataset.

It is clear from this literature that knowledge graph embeddings are capable of encoding the similarity between concepts, but the work of Hubert et al. and Jain et al. outline the fact that this performance can be highly variable, that it therefore needs to be thoroughly evaluated on a per-dataset, a per-model, and a per-class basis, and that the traditional rank-based metrics are probably not the most meaningful way of assessing this.

4.1.3 *Comparing complex concepts and sequences*

Many of the approaches used to learn word embeddings are in fact *compositional*: that is, a phrase or sentence can be represented by a simple combination of the embeddings of the individual words that make it up; often by either addition or multiplication (Mikolov et al. 2013a; Mitchell & Lapata 2008). L. White et al. (2015) establish that summing or averaging word embeddings to create sentence embeddings can be just as effective as dedicated approaches.[5] This approach to representing complex concepts as combinations of their individual components has been widely adopted: Zou et al. (2013) represent phrases as averages of their word vectors, and Nalisnick et al. (2016) create embeddings of entire documents from a normalised mean of their constituent word embeddings.

5 Note that summed and averaged vectors are effectively equivalent, since the widely used cosine distance considers only a vector's direction, and not its magnitude.

This is not, however, the only approach to generating representations of sequences of words. Incitti et al.'s literature review (2023) identifies several other methods, ranging from the simplest techniques that count word occurrence, to methods based on machine learning. What is noticeable, however, is that many of the more complex methods still use the idea of compositionality: the paragraph vector or "doc2vec" approach, for example, generates its representations by creating a representation of an entire paragraph as well as representations for each of the individual words, and then concatenating or averaging them together (Le & Mikolov 2014).

These examples demonstrate that a great number of embedding approaches for both words and graphs incorporate some notion of compositionality. Whilst this is not universal — where it does apply, it is designed in in some cases and emerges naturally in others — it does mean that in many cases, the meaning of a complex concept can be defined by a combination of the meanings of its component concepts. This simple principle is a powerful

one, but it is not without its flaws. The resulting sentences are represented in the same embedding space as the individual words, which reduces the expressive power of these approaches (Gupta et al. 2020), and most importantly these approaches are typically not sensitive to word order, meaning that two sentences consisting of the same words in a different sequence will be represented by the same embedding. This is problematic when trying to represent processes and histories of event, where order definitely does matter.

Knowledge graph embeddings operate on many of the same principles as word embeddings, so it follows that these approaches might also be effective for KGEs. Carvalho et al. (2023), in their work on hospital readmission prediction, represent a patient as a concatenated vector of all the ontology classes describing their features. Outside of this example however, the compositionality of KGE-based representations has not been widely researched (Bertolini 2023).

As identified in the earlier literature review, string metrics (Williams et al. 2014), and dynamic programming algorithms (Aspland et al. 2021) can effectively provide a distance score between two sequences of symbols, but these naturally forgo the benefits of semantic knowledge as encoded by word or knowledge graph embedding approaches. There are also plenty of algorithms from the timeseries literature that compare sequences of continuous values. such as the dynamic time warping (DTW) algorithm (Berndt & Clifford 1994). Whilst more recent methods exist (Middlehurst et al. 2024), DTW provides an effective baseline for quantifying timeseries similarity, and in particular has been widely adopted in the medical literature to cluster, sequences of continuous variables over time, including risk scores (Hebbrecht et al. 2020; Mesbah et al. 2024), vital signs (Bhavani et al. 2023), and biomarker values (Burke et al. 2022).

The work of Giannoula et al. (2018) is particularly relevant, investigating patterns of disease progression by using DTW to compare ordered vectors of diagnoses. However, this approach is limited in that it calculates distances based on the numeric value of diseases' ICD-9 codes (specifically the squared distances of the code values), a distance measure with no inherent semantic meaning. This has been extended to consider distance between codes based on their proximity in the ontology graph structure, or their binary presence or absence in the same hand-defined category (Giannoula et al. 2021, 2024), but these remain relatively simplistic measures of similarity.

This chapter therefore proposes a new method which combines the benefits of semantic similarity from knowledge graph embeddings with sequence comparison, ordering and timing from dynamic time warping. By treating a

patient's pathway as a sequence of ontology concepts, an embedding can be used to transform this sequence of symbols into a continuous space, and this sequence of embeddings can be effectively treated as a timeseries, and compared using corresponding approaches.

4.2 *Learning representations of medical concepts*

As a prerequisite to creating representations of a patient's pathway, I first consider the process of learning representations of the individual concepts that comprise it. In this section, I evaluate several KGE models by creating embeddings of the concepts and relationships in the SNOMED ontology. As well as investigating their performance on the conventional link prediction metrics, I specifically evaluate how well they represent semantic similarity in terms of placing similar concepts close to each other in embedding space.

4.2.1 *Methods*

The method used for training KGE models was based on that described by Chang et al. (2020) Models were trained on the April 2024 release of the SNOMED CT UK Clinical Edition (`SNOMEDCT2_38.0.0_20240410000001`). The ontology was filtered down to a subset of the most informative semantic groups, but rather than use the Unified Medical Language System (UMLS) system's semantic categories, this was done based on the equivalent SNOMED top-level concepts: this simplifies the process, removing the need to introduce another layer of ontology, and ensures a clear provenance of training data, guaranteeing that the embedding is influenced only by information present in the specified release of SNOMED. The selected groups were *Body structure*, *Artifact*, *Clinical finding*, *Event*, *Procedure*, *Pharmaceutical/biologic product*, *Device*, and *Substance*; any concept that is a direct or transitive descendant of any of these top-level concepts is included in training.

The graph is treated as bidirectional, and triples describing reciprocal relations are included. Embeddings were trained, on an NVIDIA Tesla V100 graphics processing unit with 16GB of memory, using the `PyKEEN` library (Ali et al. 2021) in Python. In all experiments, the data was randomly split into training and testing sets at a rate of 95%:5%, following the example of Chang et al. (2020). Where an embedding model was described in Chang et al.'s benchmarking paper, their recommended hyperparameters were used to make our results comparable, although deviation from their results is expected given that a more recent and UK-specific release of SNOMED is used, and slightly different inclusion criteria are defined. Three of the most popular KGE models were compared:

TransE is the seminal knowledge graph embedding algorithm, and is widely used as a benchmark for comparison (Bordes et al. 2013). For a given triple (h, r, t), TransE treats r as a translation vector that approximately transforms h's vector into t's; i.e. $\mathbf{h} + \mathbf{r} \approx \mathbf{t}$. Therefore, the scoring function is simply the squared Euclidean distance between $\mathbf{h} + \mathbf{r}$ and $\mathbf{t}$: [6]

$$f_r(h, t) \;=\; -\| \mathbf{h} + \mathbf{r} - \mathbf{t} \|_2^2 \tag{4.1}$$

A smaller distance between $h + r$ and t results in a higher score, indicating that a triple *holds*, and the fact it represents is more likely to be true according to the source knowledge graph.

TransE achieves MRR scores of around 0.4–0.5 on the FB15K and WN18 datasets, two classic benchmarks, but these drop to 0.2–0.3 for the FB15K-237 and WN18RR datasets, which introduce more complex relation patterns (Sun et al. 2019). On SNOMED specifically, Chang et al. (2020) achieve an MRR score of 0.346.

TransE has several limitations: it struggles to embed anything more complex than a one–to–one relationship. Since $h + r$ must $\approx t$ for multiple ts, it is inevitable that several distinct entities will begin to receive very similar embedding values ($t_1 \approx t_2 \approx ... \approx t_n$:). This has given rise to many further embedding models which aim to resolve these and other issues.

DistMult models entities as vectors and relations as matrices (B. Yang et al. 2015). For each relation, a matrix $\mathbf{W}_r \in \mathbb{R}^{d \times d}$ contains weights $w_{i,j}$ that capture the amount of interaction between the ith latent factor of the head entity and the jth latent factor of the tail entity. These are restricted to diagonal matrices, a step that reduces the number of parameters and improves performance compared to its closest ancestor model RESCAL (Nickel et al. 2011, 2012), but which prevents it from being able to model anti-symmetric relations.

$$f(h, r, t) \;=\; \mathbf{h}^\top \mathbf{W}_r \mathbf{t} \;=\; \sum_{i=1}^{d} \mathbf{h}_i \cdot \mathrm{diag}\,(\mathbf{W}_r)_i \cdot \mathbf{t}_i \tag{4.2}$$

DistMult achieves very high performance (almost 0.8) on the FB15K and WN18 benchmarks, but these drop dramatically to 0.2 and 0.4 on the more complex versions. On SNOMED, it achieves a more modest 0.420.

RotatE models relations as rotations from head to tail entities in a complex space, and is able to model symmetry, antisymmetry, inversion and composition (Sun et al. 2019).

$$f(h, r, t) \;=\; -\| \mathbf{h} \odot \mathbf{r} - \mathbf{t} \|_1 \tag{4.3}$$

6 This and subsequent equations use the following conventions:

s	scalar quantities
$\mathbf{v}$	vectors
$\mathbf{M}$	matrices
$\mathbf{M}^\top$	transposed matrix
$\mathbf{M} \odot \mathbf{N}$	Hadamard (element-wise) product of matrices
$\| x \|_1$	L^1 (Manhattan) norm of x
$\| x \|_2$	L^2 (Euclidean) norm of x

RotatE follows a similar pattern to DistMult, achieving very high scores (0.8–0.95) on classic benchmarks and lower scores (0.3–0.5) on their more complex counterparts; on SNOMED, this performance is towards the lower end (0.3).

4.2.2 *Evaluation*

Each model is evaluated in three different ways: according to conventional link prediction metrics, according to their encoding of semantic similarity as compared to the OPCS terminology, and according to their encoding of semantic similarity as measured by the original graph's neighbourhoods.

Link prediction

Each model is first evaluated on the task of link prediction, as is conventional in the KGE literature. For every triple in the test set, Q, a corresponding set of false triples is created by randomly swapping either the triple's head, relation, or tail with a random entity. Each triple's plausibility score is calculated, according to the particular model's scoring function, and the original true triple's rank amongst these scores is stored. From this, two statistics are derived:

Mean reciprocal rank the mean of all true triples' predicted ranks, normalised to range from 0 (worst performance) to 1 (best):

$$\text{MRR} = \frac{1}{|Q|} \sum_{t \in Q} \frac{1}{\text{rank}(t)} \tag{4.4}$$

Hits@k the proportion of occasions where the true triple was ranked within the top k scoring triples:

$$\text{hits@}k = \frac{|\{\, t \in Q \mid \text{rank}(t) \geq k \,\}|}{|Q|} \tag{4.5}$$

These statistics are widely used in the evaluation of KGE models, and therefore provide a point of comparison (Ali et al. 2022).

Corrupting triples in this way can, by chance, occasionally produce true triples which would then be scored as if they were false. For this reason, a common adjustment is to automatically discard from the list of corrupted triples any triple that exists in any training, validation or test set, other than the triple under evaluation (Bordes et al. 2013). When working with the SNOMED ontology, Chang et al. (2020) extend this further, recommending that all transitive "is a" relations should be calculated and also excluded. This approach is used here, in order to ensure that the evaluation results

are comparable with one of the major benchmarking papers for SNOMED embeddings.

Ontological structure

To evaluate how well the learned embeddings of SNOMED concepts represent notions of semantic similarity, I adapt the method described by Fu et al. (2023) for evaluating embeddings of ICD concepts. This method was originally proposed for evaluating embeddings of ICD diagnosis codes, but the principles can also be applied to OPCS procedure codes. Each code in the OPCS terminology is mapped to its corresponding SNOMED concepts, via the mappings in the OMOP Common Data Model (Hallinan et al. 2024), and these SNOMED concepts are mapped to their learned embeddings. Since an OPCS code can map to multiple SNOMED concepts, each OPCS code is represented by an $n \times d$ matrix, where n is the number of SNOMED codes mapped to that concept, and d is the embedding dimensionality.

The similarity of two concepts can then be described by measuring the distance between them in the embedding space. Conventionally, this is done using the cosine distance:

$$\cos(\mathbf{x}, \mathbf{y}) = 1 - \frac{\mathbf{x} \cdot \mathbf{y}}{\| \mathbf{x} \|_2 \| \mathbf{y} \|_2} \tag{4.6}$$

Fu et al., however, instead use a modified version of the RV coefficient. The classic RV coefficient (Robert & Escoufier 1976), for two matrices $\mathbf{X}$ and $\mathbf{Y}$ of dimensions $n \times p$ and $n \times q$ respectively is calculated as:

$$RV(\mathbf{X}, \mathbf{Y}) = \frac{\sum_{i=1}^{p} \sum_{j=1}^{q} r^2\left(\mathbf{x}_i, \mathbf{y}_j\right)}{\sqrt{\sum_{i,j=1}^{p} r^2\left(\mathbf{x}_i, \mathbf{x}_j\right) \sum_{i,j=1}^{q} r^2\left(\mathbf{y}_i, \mathbf{y}_j\right)}} \tag{4.7}$$

where $r^2(\mathbf{a}, \mathbf{b})$ represents the squared Pearson correlation between two vectors. Mayer et al. (2011) note that for high-dimensional datasets, RV can become misleadingly high, and recommend using a modified r^2 that adjusts for the dimensionality n:

$$r^2_{adj}(x, y) = \frac{n-1}{n-2}\left(1 - r^2(x, y)\right) \tag{4.8}$$

The RV coefficient is effectively a multivariate generalisation of the Pearson correlation, which is itself equivalent to the squared cosine distance when the data is centred (i.e. the mean is zero). Therefore, the RV coefficient is intuitively very similar to the cosine distance, but it importantly allows for the comparison of two matrices with different numbers of columns; a useful property in our situation where an OPCS code may map to multiple

SNOMED concepts, and therefore multiple embedding vectors. The RV coefficient therefore provides a more appropriate metric to compare each OPCS code's $n \times d$ matrix.

To measure the extent to which the embedding distances reflect structural information about concepts, the average RV distance between pairs of OPCS codes from the same chapter is compared to the average RV distance between pairs of codes in two different chapters. The RV distance metric is also benchmarked against an alternative metric, a simple tree-based distance which encodes the hierarchical structure of the SNOMED ontology. In this measure, the distance between two codes A and B is always 1 if they are from different chapters, or otherwise the difference between their numeric components multiplied by 0.01:

$$\text{dist}_{\text{tree}}(A, B) = \begin{cases} 1 & \text{if } A_{chap} \neq B_{chap} \\ (A_{num} - B_{num}) \times 0.01 & \text{otherwise} \end{cases} \tag{4.9}$$

Fu et al. also consider the frequency of co-occurence between codes, measured using the Jaccard distance, in their analysis. However, the vast majority of codes never co-occur in their analysis, heavily skewing the Jaccard values towards 1 (i.e. maximally distant). Additionally, the data excerpt available for this analysis only contained information on colorectal cancer patients, meaning that any co-occurrence data would not be representative of that concept in general. These factors severely limit the usefulness of this metric; for these reasons, this step is omitted from this analysis.

Graph structure

To measure how well an embedding preserves the original graph's structure, each concept's set of closest neighbours according to the original graph structure is compared to the set of closest neighbours according to the learned embedding, using the method outlined by Hubert et al. (2024).

For each concept A, its "triple neighbourhood" $N(A)$ is defined as the set of all triples involving A as either the head or tail:

$$N(A) = \{ (A, r, t) \cup (h, r, A) \} \tag{4.10}$$

After replacing A with a dummy identifier that remains constant across all values of A, the similarity between two concepts is defined as the Jaccard index between their respective triple neighbourhoods:

$$J(c_1, c_2) = \frac{| N(c_1) \cap N(c_2) |}{| N(c_1) \cup N(c_2) |} \tag{4.11}$$

Model	Link prediction			Graph neighbourhood		OPCS
	MRR	H@1	H@10	R1@10	R1@100	Correlation
TransE	0.407	0.334	0.542	0.150	0.108	0.071
DistMult	0.360	0.274	0.523	0.226	0.146	0.198
RotatE	0.727	0.633	0.840	0.266	0.193	0.184

Table 4.1. Performance of KGE models measured by according to link prediction metrics, neighbourhood similarity to the original graph (Hubert et al. 2024), and correlation with the OPCS codesystem's structure (Fu et al. 2023)

Each concept's k closest neighbours according to the Jaccard similarity is then compared with the same concept's k closest neighbours according to their cosine distance in the embedding space. The similarity between these two concept neighbourhoods is quantified with rank-biased overlap (RBO). For the two ranked lists of concepts S and T, their agreement to a given depth d is defined as the intersection of the first d items on both lists, weighted by d:

$$A_{S,T,d} = \frac{|\, S_{1:d} \cap T_{1:d} \,|}{d} \tag{4.12}$$

The overall RBO for the two lists of length k can then be calculated

$$\mathrm{RBO}\,(S, T, k) = \frac{1}{k} \sum_{d=1}^{k} A_{S,T,d} \tag{4.13}$$

RBO therefore measures the overlap between two ordered lists of concepts, whilst also being sensitive to their ordering, and giving a higher weighting to matches that occur nearer to the top of the list.

4.2.3 *Results*

In terms of MRR, TransE's performance exceeded its comparable benchmark, and DistMult performed slightly worse. RotatE, however, significantly exceeded the previous benchmark, also achieving significantly higher hits@k performance.

In terms of the RBO scores, RotatE was again the best model, but the improvement is not as dramatic. This indicates that whilst there does appear to be a relationship between link prediction performance and the capturing of graph neighbourhood structure, this relationship does not necessarily scale linearly. In fact, RotatE exhibits a lesser correlation with OPCS tree distance than DistMult, although this could be due a result of encoding other similarities not necessarily obvious from OPCS structure. For each embedding, the full evaluation of OPCS structure is presented in Appendix A. TransE exhibits a generally low distance between every concept, indicating relatively poor separation of concepts by chapter. DistMult provides a substantially improved representation, with a clear separation in the distri-

bution of within-chapter (lower quartile $Q_1 = 0.35$, median $M = 0.47$, upper quartile $Q_3 = 0.5$) and cross-chapter ($Q_1 = 0.61$, $M = 0.75$, $Q_3 = 0.85$) distances. There is a correlation between embedding distance and tree structure, although some individual OPCS chapters display negative correlations suggesting varying effectiveness on different topics. RotatE produces a slightly weaker correlation between embedding and tree distances, but also creates a very pronounced separation between within- and between-chapter concepts ($Q_1 = 0.89$, $M = 0.90$, $Q_3 = 0.91$ versus $Q_1 = 0.97$, $M = 0.98$, $Q_3 = 0.98$).

4.3 *Constructing and comparing representations of patient pathways*

The previous section used KGE models to generate embeddings of biomedical concepts from the SNOMED ontology. This section proposes a method for combining these individual concept representations into a representation of a patient's entire pathway.

4.3.1 *Methods*

As summarised in Section 4.1.3, there are several possible methods for combining individual embeddings into representations of more complex concepts. In this section I outline the data used, the proposed method for creating patient representations, and several existing methods used for comparison purposes. The method described here supports two tasks: it allows a user to combine an EHR dataset with a pre-trained set of concept embeddings, i.e. the output of a KGE model, and generate vector representations of the patients' pathways; and it allows the distance between these pathways to be described with a numerical value.

Data

The following experiments used the OUH dataset previously described in Section 3.2.2. Two cohorts were identified, consisting of patients with a recorded colon (`C18`) or rectal (`C20`) cancer diagnosis in the cancer management system between December 2012 and February 2024 inclusive ($n = 7,392$). From this dataset, each patient's history of clinical events is extracted and represented as an ordered sequence of events $P_i = \{e_1, ..., e_n\}$.

Each event e_n can be one of three types. Procedures are any events represented by an OPCS code listed in the CORECT-R codelist (*CORECT-R Data Coding* 2020). These are directly represented as the relevant OPCS code. Chemotherapy is represented as a "chemotherapy" marker. Chemotherapy is delivered as a *course*, which is divided into *cycles*: here, one "chemotherapy"

event indicates a cycle, to ensure that information around course length and timing is implicitly included.

Radiotherapy is likewise represented as a "radiotherapy" marker. However, counting the number of radiotherapy OPCS codes in the main procedures table, the number of entries in the radiotherapy table, and the number of outpatient attendances with "radiotherapy" as the given reason gives three very different results for the number of patients undergoing radiotherapy. I used the second option, data from the radiotherapy table, as it gives the largest number. It however only summarises radiotherapy regimens, and does not describe the number of attendances: the pattern of actual radiotherapy delivery was therefore estimated by taking the recorded number of fractions delivered, and assuming that they were delivered on every weekday from the regimen's start date. It is usual practice to deliver fractions daily in this way (Royal College of Radiologists 2019), and in the actual data the number of fractions delivered (plus weekends) was usually approximately equal to the number of days between between the recorded start radiotherapy start date and end date (±10 days in 75% of cases).

Benchmark methods

As well as the proposed method, outlined in the next section, I also use several existing distance metrics to measure the differences between patients, for comparison purposes. Firstly, I consider very basic compositions of the individual embeddings:

- Product, mean, and mean plus variance, in which the patient is a simple combination of each of the individual event embeddings, and similarity is measured as the cosine distance between them.

Secondly, I consider sequence-based methods that represent the patient as an ordered list of their OPCS procedure codes, and compare these ordered lists without considering semantic distance or timing:

- The longest common subsequence (LCS) is, given two sequences, the longest sequence common to both (Maier 1978). The LCS distance is therefore defined as $1 - \text{LCS}$, such that longer sequences in common means a smaller distance value.
- The Damerau-Levenshtein distance (DL) is the number of operations — either insertions, deletions, or substitutions of a single character, or transposition of two adjacent characters — required to transform one sequence into another (Damerau 1964; Levenshtein 1966).
- The Needleman-Wunsch algorithm (NW) was developed in bioinformatics for calculating alignments between DNA sequences. It is similar to the Levenshtein distance, assigning a score based on the number

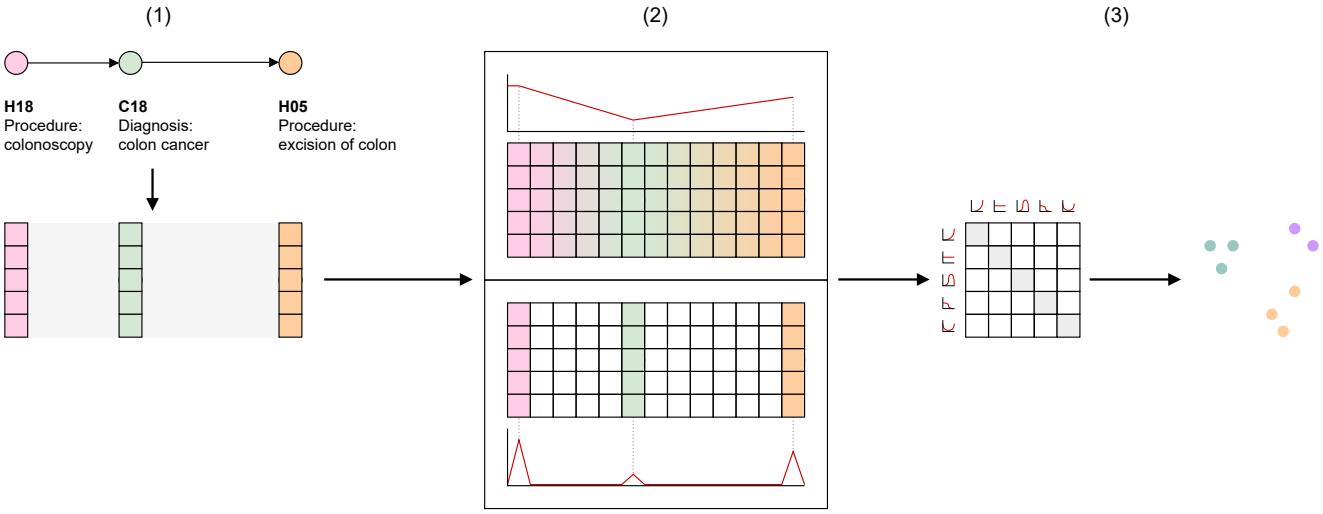

Figure 4.1. Proposed method for constructing representations of patient pathways

of operations needed to align a sequence, but allows the penalties for symbol matches, swaps, and gaps to be specified by the user (classically 1, −1, and −1 respectively).

Each of these metrics has been at some point used to compare sequences of clinical events (Aspland et al. 2021; Williams et al. 2014).

Embedding-based dynamic time warping

The proposed method, *embedding-based dynamic time warping* (E-DTW), is summarised in Figure 4.1: patient timeseries are created (1), then converted into timeseries of embeddings (2), and the distance between them measured using DTW.

A patient's timeseries T_n is a sequence of vectors $\{v_1, ..., v_n\}$ in which each vector represents a 24-hour time window. For each v_i, if the patient has an event in the corresponding time window, then the vector v_i is the embedding representing that event, according to the RotatE KGE model described in the previous section. For events encoded with OPCS codes, that vector is found by mapping the OPCS code to its closest equivalent SNOMED concept using the OMOP CDM mapping. For events without a direct mapping, an equivalent SNOMED code is manually assigned.[7] Where multiple events occur in the same time window, the later event is moved to the next window. The gaps between these events can be handled in one of two ways: either the value is set to a vector of zeroes, or the empty spaces are filled using a linear interpolation, such that the empty spaces form a gradient between two known points.

The difference between two patient timeseries is measured using dynamic time warping (DTW), specifically Meert et al.'s implementation (2020). DTW has been used in a healthcare context to compare both sequences of

7 The SNOMED concept `367336001 | Chemotherapy (procedure) |` was used for chemotherapy and `1287742003 | Radiotherapy (procedure) |` for radiotherapy.

continuous variables, and sequences of discrete procedure codes, but it can also be applied to timeseries of multidimensional vectors. First, a distance matrix $\mathbf{M}(a, b)$ is defined between two sequences a and b, describing the distances between each pair of points in a and b (each $\mathbf{M}_{i,j} = (a_i - b_j)^2$). In our case, this is the cosine distance between each embedding vector. A warping path P is defined as any sequence of points that describe a particular traversal through $\mathbf{M}$, and a warping path's distance is simply the sum of each of the points:

$$\text{dist}(a, b) = \sum_{i=1}^{s} p_i \tag{4.14}$$

The best point-to-point alignment of the two sequences is therefore the path through $\mathbf{M}$ that minimises this total distance (Bagnall et al. 2017). This is illustrated in Figure 4.2: the heatmap shows the distance between each pair of points on two timeseries, and the red line shows the traversal through the matrix that minimises the total distance; any deviation from a perfect diagonal means that an edit is required to align the two sequences.[8]

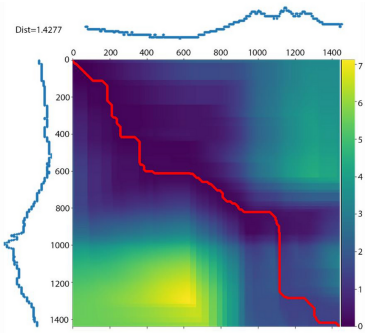

Figure 4.2. An example of a DTW matrix (Li et al. 2021)

8 Mueen & Keogh (2016) provide an extensive tutorial on dynamic time warping.

This shortest distance effectively summarises the total movement required to align the two timeseries, and can therefore be used as a distance measure. Additionally, a user-defined weighting means any non-diagonal movement in the warping path can be penalised (Jeong et al. 2011): this means that the final distance is increased as more warping is required to align the sequences, quantifying not just differences in the timeseries' peaks, but also in their spacing.

There are several modifications that can be made to this basic algorithm. The embeddings in the first step can be constructed from the original high-dimensional embeddings, or they can be reduced to a smaller number of dimensions using principle component analysis (PCA). The gaps between events can be padded using zeroes, or by interpolating values. The DTW algorithm can use a warping penalty of 0 or 1. Combined, these parameters allow for eight different variations on our embedding DTW algorithm; all eight of these are tested in this evaluation.

The proposed method therefore involves taking a patient's clinical pathway, representing it as a sequence of embeddings, and then treating this sequence of embeddings as a multidimensional timeseries or a sequence of points in the embedding space, allowing DTW to be applied to quantify the difference in both time and space.

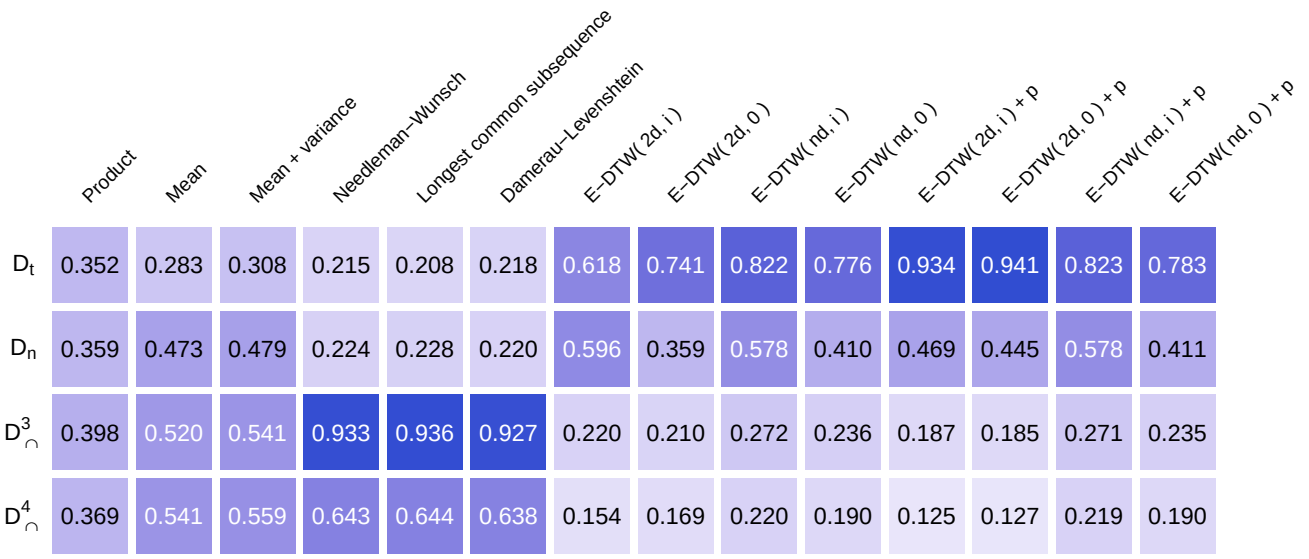

Figure 4.3. Pearson correlation between different pathway distance metrics

4.3.2 *Results*

Figures 4.4 and 4.5 show the relationship between the described pathway distance metrics and several different properties of each pathway:

- D_t is defined as the absolute difference in two different pathways' timespans (the time elapsed between the first and last event, in days)
- D_n is defined as the absolute difference in the number of events in two different pathways
- $D^3_\cap$ is defined as the number of three-character OPCS codes shared by both pathways
- $D^4_\cap$ is defined as the number of four-character OPCS codes shared by both pathways

Comparisons are made against eight different variants of the E-DTW algorithm. For brevity these are referred to using the notation E-DTW (d, f) $[+p]$, where d represents the dimensionality of the embeddings used ($2d$ or nd), f represents the padding strategy used to fill empty space (zeroes, 0, or interpolated, i), and the addition of $+p$ indicates a warping penalty of 1.

Figure 4.3 summarises each of these relationships using the Pearson correlation coefficient (ρ). Overall, the E-DTW approach correlated more strongly with D_t and D_n, whilst the simple and sequence-based metrics correlated more strongly with $D^3_\cap$ and $D^4_\cap$ The basic metrics (product, mean, mean+variance) all displayed weak correlation with D_t (ρ = 0.352, 0.283, 0.308). This is largely expected as they summarise the semantic content of a patient's embeddings without reference to timing. Sequence-based metrics (NW, LCS, DL) displayed a similar pattern (ρ = 0.215, 0.208, 0.218), which is

similarly expected as these methods treat pathways as a list of events with no timing information.

By contrast, every variation of E-DTW displayed a moderate to strong correlation with D_t: in particular, it is remarkable that the variants using two-dimensional embeddings and a warping penalty (E-DTW $(2d, *) + p$) showed an extremely strong association (ρ = 0.934, 0.941). Figure 4.4, which show the distribution of individual distance pairs, allows this difference to be visualised. Every other zero-padded variant was clearly divided into two regions: one set of distances with a good correlation, plus an extra horizontal band of distances with close proximity to each other, but no correlation against timespan. Every other interpolated variant displayed a roughly evenly distributed "cloud" around the main strong trend.

Overall, variants using two-dimensional embeddings produce a much more defined correlation with D_t, but at the cost of creating a separate cloud of unassigned points; interpolating embeddings to fill gaps produces a generally fuzzier main arc, but leave far fewer dramatic outliers. In every case, the introduction of a warping penalty increased the level of correlation with D_t compared to its non-penalty equivalent. This makes sense: warping penalties explicitly penalise the distance score and increase it for every edit that has to be made to align the two sequences. The remarkable pattern seen in E-DTW $(2d, *) + p$ variants is likely because reducing the embedding dimensionality inevitably reduces the amount of semantic content captured: these versions therefore have the minimal semantic content (due to the dimensionality reduction) and the maximum temporal content (due to the warping penalty), and therefore weight temporal information significantly higher.

The basic and sequence-based metrics were only weakly correlated with D_n. Introducing penalties did not lead to strong change in correlation compared to those versions without penalties; a greater number of individual events does not necessarily mean a longer timespan. The distribution of values for these resembled the pattern seen in D_t, with a weaker main correlation and a much larger patch of outliers.

Conversely, almost all of the basic and sequence-based measures showed a moderate to strong correlation with $D^3_\cap$ and $D^4_\cap$. The product metric was the outlier, perhaps because any difference in the embeddings' semantic content was outweighed by the multiplicative factor. E-DTW did not perform as strongly, and the $+p$ variants where typically weaker since, for the reasons previously outlined, they prioritise timing. The strongest correlations with number of shared events were found in the full-dimensional, interpolated variants, which maximised semantic content.

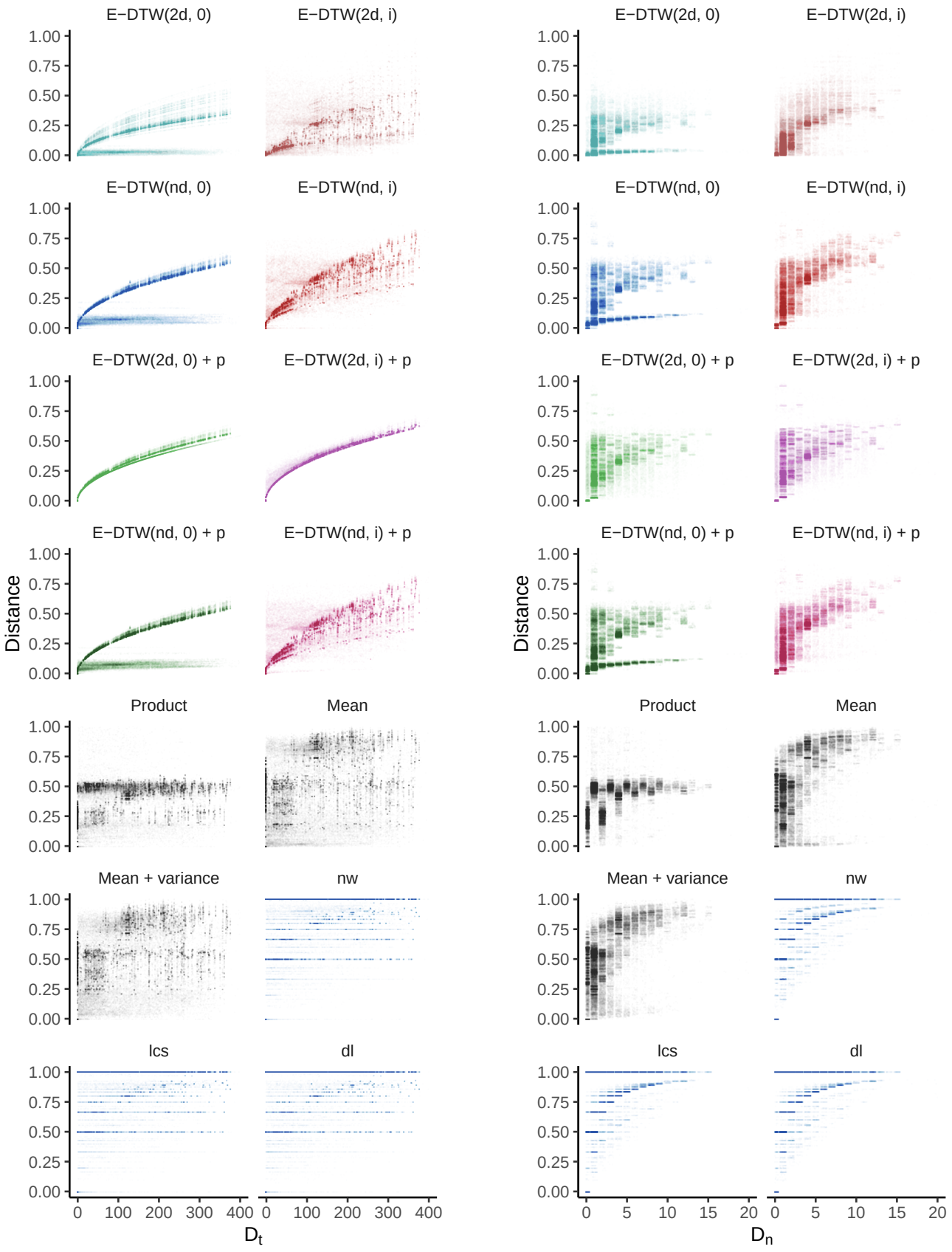

Figure 4.4. Distributions of pairwise pathway-pathway distances, according to various distance metrics (Y) against D_t (X, left) and D_n (X, right)

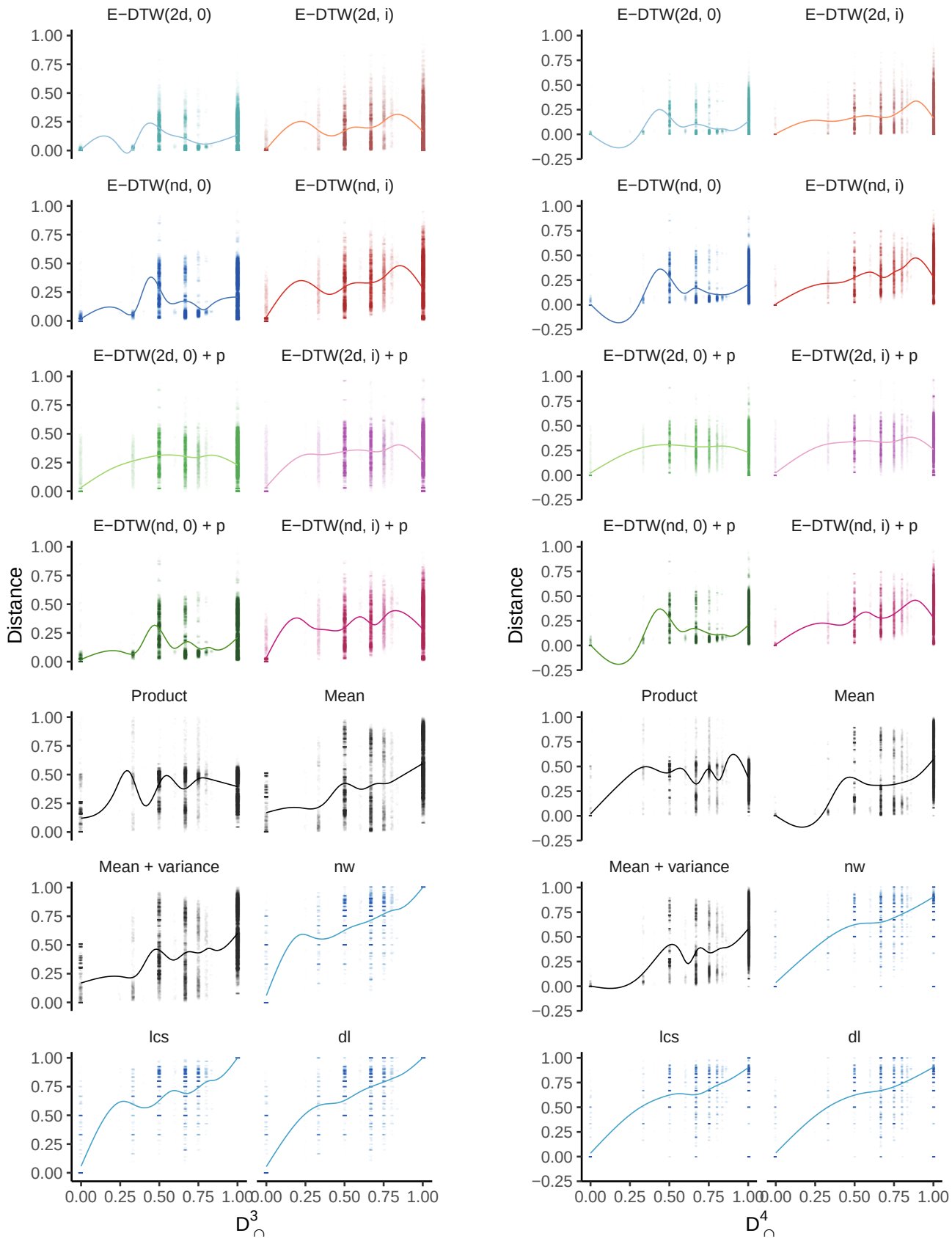

Figure 4.5. Distributions of pairwise pathway-pathway distances, according to various distance metrics (Y) against $D^3_\cap$ (X, left) and $D^4_\cap$ (X, right)

What is noticeable from Figure 4.5 is that in almost every distribution, but especially those for E-DTW, there exist a large number of pathways where $D_\cap = 1$ (i.e. maximally different with no shared events at all), but which occupied a range of embedding distance values, creating a distinctive band down the right hand side of every plot. The exception to this was the sequence-based metrics, which agreed with $D_\cap = 1$ on values of 1 far more often. These differences suggest that the E-DTW metric can consider two pathways to be similar even if the codes are not strictly the same, or from the same chapter — it has a less binary view of concept similarity. Similarly, it allows two sequences with many events in common to still be considered different if they have sufficiently different timing.

In addition to comparing distance metrics against properties of each pathway, the results of each were visualised by randomly sampling 32 patients of different pathway lengths, and retrieved their twenty nearest neighbours according to each distance metric. These are displayed in full in Appendix B. Overall, product, mean, and mean + variance all tended to retrieve pathways that were thematically similar, i.e. containing a similar set of events, but which had very different temporal patterns. LCS, NW, and DL were better able to retrieve temporal motifs, but did sometimes retrieve pathway exhibiting the same timing patterns but with different events and orders. The E-DTW metrics however provided the best results, typically encoding both sequential patterns and semantic similarity. E-DTW $(2d, 0)$ was noticeably good at encoding sequences: for example, the very bottom-right pathway correctly retrieved neighbours with the sequence of surgery then, chemotherapy, but which tended to be of more varying length. E-DTW $(2d, i)$ provided a slight improvement in terms of length. E-DTW $(nd, 0)$ and E-DTW (nd, i) had a similar relationship to each other.

E-DTW $(2d, 0) + p$ placed a stronger emphasis on matching length, perhaps too rigidly: it matched pathways where they fitted the same temporal patterns, even if the events were totally different. E-DTW $(2d, i) + p$ was similar, but better at matching similar events. E-DTW $(nd, 0) + p$ and E-DTW $(nd, i) + p$ were both excellent at matching sequences of events, the key differences being that the interpolated version appeared slightly more flexible on length matching.

In summary, simple composition of embeddings was able to represent the semantic similarity between pathway content, and classic sequence metrics were able to represent temporal patterns, but the proposed E-DTW combined the benefits of both. In terms of fine-tuning E-DTW performance, multidimensional embeddings were — at the cost of increased computational requirements — largely more effective at encoding semantic content than

two-dimensional representations. Filling gaps in pathways with interpolated values representing the gradient between two concepts appeared to produce a slight improvement in the algorithm's ability to match sequences by length. The introduction of a warping penalty improved the level of temporal alignment, but for two-dimensional embeddings, this adjustment was overwhelming and overrode any semantic similarity, whilst when combined with higher-dimensional embeddings it provided a useful complement.

The "best" E-DTW configuration therefore depends on the context in which it is being used. In scenarios where emphasis is placed on temporal similarity, the E-DTW $(2d, i) + p$ configuration provides an excellent matching of temporal patterns, whilst reducing the confusion of semantically distinct events associated with the E-DTW $(2d, 0) + p$ version. Where a broad semantic similarity between events is desired, either E-DTW $(2d, 0)$ or E-DTW $(2d, i)$ suffice. In most scenarios, where a balance between the two is advantageous, E-DTW $(nd, i) + p$ provides an appropriate combination, effectively encoding both elements whilst avoiding overly strict temporal matching.

4.4 *Applications in cancer care*

Having developed the E-DTW method, and confirmed that it encodes meaningful semantic differences between patient pathways, I now consider its usefulness in answering real research questions. I apply the E-DTW measure to a specific case study, in order to evaluate its usefulness and also to identify further considerations that arise from its usage in answering research questions.

The use of radiotherapy in addition to surgery to treat rectal cancer has been proven to reduce the chances of local recurrence. However, studies have at the same time suggested that it increases the risk of longer-term complications, and has little effect on overall survival for moderate-risk disease. This somewhat mixed picture means that there is no agreed standard for the use of radiotherapy in conjunction with surgery, and patterns of usage vary widely, both internationally and even between different providers within the UK.

E. J. A. Morris et al. (2016) studied 9,201 individuals who were diagnosed with rectal cancer (ICD-10 code C20) and underwent a major resection in England over an 18 month period. In this study, individuals were allocated to one of five groups, based on the radiotherapy regimen they experienced:

- *No radiotherapy* (NRT), where there was no record of any radiotherapy

- *Short-course radiotherapy and immediate surgery* (SCRT-I), where a patient attended a radiotherapy centre five times before surgery, and where the time between start of radiotherapy and surgery was less than 35 days
- *Short-course radiotherapy and delayed surgery* (SCRT-D), where a patient attended a radiotherapy centre five times before surgery, but the time between the start of radiotherapy and surgery was greater than 35 days
- *Long-course chemoradiotherapy* (LCCRT), where a patient either attended for radiotherapy 25, 28, or 30 times; or had multiple individual radiotherapy records with at least one record describing at least 10 attendances and an overall total of at least 25 attendances
- *Post-operative radiotherapy* (PORT), where a patient received any radiotherapy up to a year after surgery
- *Other radiotherapy* (ORT), where a patient received radiotherapy not meeting any of the above criteria

Across the whole study population, the largest group was NRT (50.7%), followed by LCCRT (29.6%), SCRT-I (12.1%), ORT (4.7%), PORT (2.3%), and SCRT-D (1.2%). However, this distribution changes significantly when broken down into geographic regions: the proportion of patients in the NRT group, for example, was as low as 22.2% and as high as 94.9%, and the proportion in the ORT group ranged from 1.1% to 29.6%. This study provides a useful benchmark for studying patterns of radiotherapy: it tells us broadly what the expected patterns of treatment would be, as defined by a domain expert, and it tells us approximately how common they should be. This section investigates the extent to which the pathway patterns discovered by our E-DTW algorithm correspond to these known groups.

4.4.1 *Methods*

In this study, a set of distances between patient pathways were derived from the E-DTW method, and the pathways were clustered to identify common treatment groups. For every patient with a `C20` diagnosis code, a timeline was created consisting of all surgery and radiotherapy events from one month before the date of diagnosis to one year after the data of diagnosis, based on the inclusion criteria used in official NHS statistics (National Disease Registration Service n.d.). A surgery was included if it was listed as a "primary procedure" by the CORECT-R codelist, incorporating the major and minor resection, bypass, stoma or stent categories.

These timelines were converted into an embedding timeseries using the method previously outlined in Section 4.3, and clustered using the Markov

cluster algorithm (van Dongen 2008). In this algorithm, the matrix of E-DTW distances is interpreted as a graph, with each nodes representing a patient, and edges between every pair of patients weighted according to their E-DTW similarity (1 − distance). This matrix is repeatedly *expanded*, i.e. squared, and *inflated*, i.e. each element raised to a specified power r and the columns scaled to sum to 1. Densely connected regions create groups of nodes which naturally contain a higher number of longer random walks; if the weightings are viewed as transition probabilities between nodes, then the expansion process boosts the probabilities of these longer walks, and penalises probabilities that connect different dense regions. Repeated, this process gradually partitions the graph into distinct regions (Figure 4.6).

The single parameter r controls inflation, which affects cluster granularity, and can be specified according to the user's needs. Additionally, I also define a threshold t, which is used to remove edges with low similarities, whilst still retaining at least one edge for each node. These two parameters mean that the clustering process can, if desired, be encouraged to leave pathways with no clear neighbours as outliers, rather than forcing them into any cluster, a desirable feature given the heterogeneity of patient pathways. The user is able to specify the readiness of the algorithm to do this to their individual needs by varying these parameters.

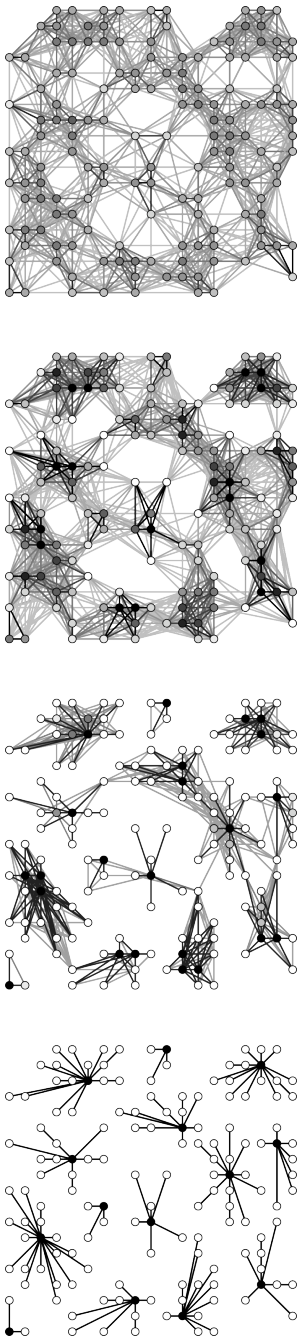

Figure 4.6. The MCL process (van Dongen 2000)

4.4.2 *Results*

712 patients with a rectal cancer diagnosis were identified, of which 347 (48.7%) were recorded as having undergone a surgery from the CORECT-R list. The clustering process was repeated, varying t and r each time, and values was selected that that produced six major clusters — a "major" cluster being defined as any cluster with ≥ 10 members — to correspond with Morris et al.'s six treatment groups. The major clusters found (where t = 78th percentile of distance scores, $r = 1.5$) were:

A No radiotherapy, and one surgery only ($n = 141$)

B No radiotherapy, two surgical operations with a short gap ($n = 43$)

C Pre-operative radiotherapy: a period of radiotherapy, followed by surgery ($n = 37$)

D No radiotherapy, two surgical operations with a longer gap ($n = 24$)

E Post-operative radiotherapy: patients received surgery, followed by a period of radiotherapy ($n = 17$)

F Post-operative radiotherapy: patients received surgery, followed by further surgery ($n = 10$)

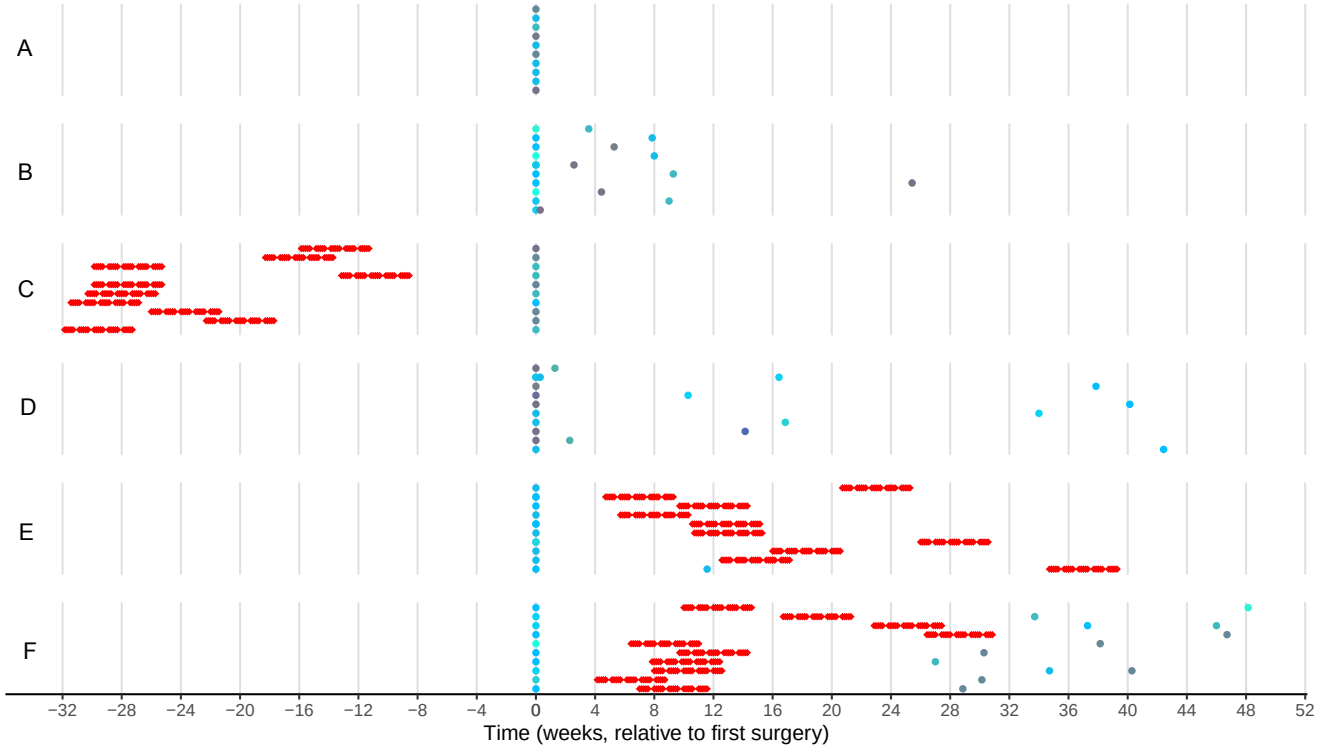

Figure 4.7. Example pathways from the major rectal cancer clusters. Ten example patients were randomly sampled from each cluster; each row represents a single patient's timeline, with symbols marking events according to the key in Figure 4.8

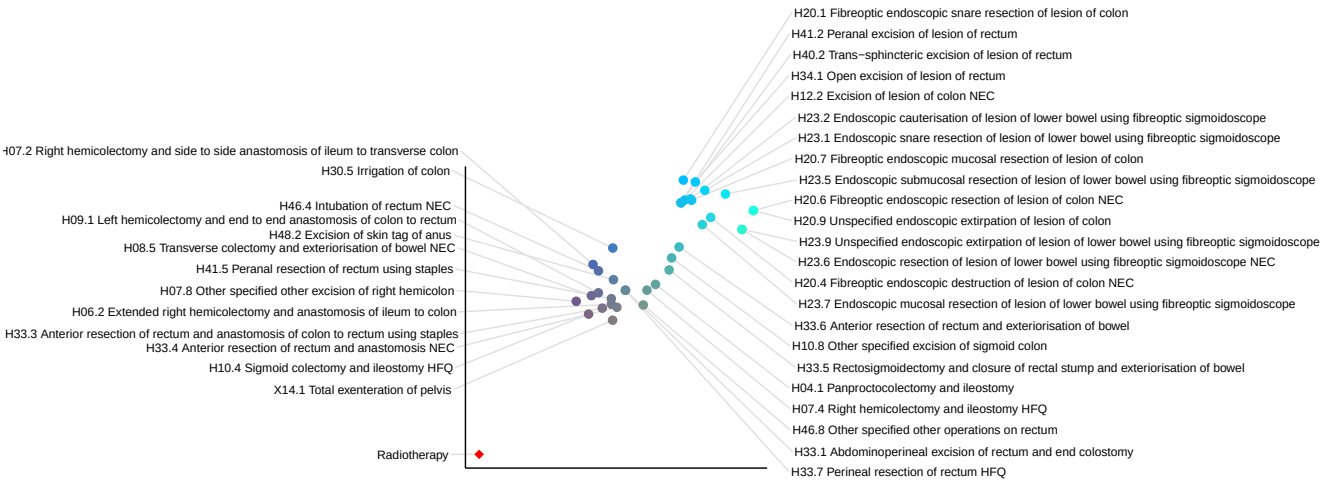

Figure 4.8. Key to OPCS concepts used in Figure 4.7

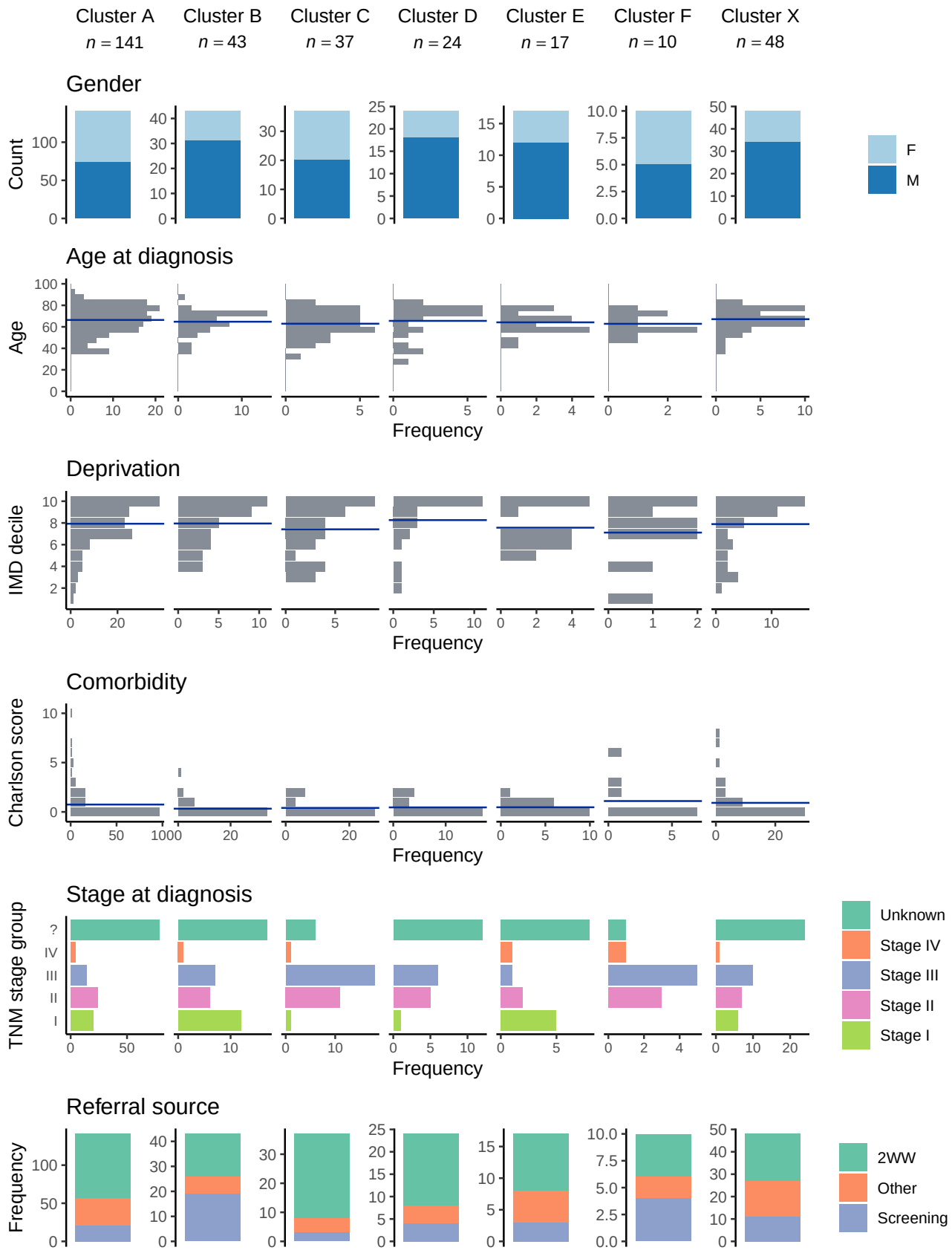

Figure 4.9. Demographic characteristics of identified rectal cancer clusters

Additionally, an artificial cluster was formed to investigate whether unusual and outlying pathways had any factors in common:

x Any patient in a cluster of nine or fewer members (total $n = 48$)

Samples from each cluster are show in Figure 4.7; Figure 4.8 shows a two-dimensional projection (via PCA) of the concepts' embeddings, and therefore a key for Figure 4.7's symbols.

The expected groupings are present here, to varying degrees: PORT is directly equivalent to cluster E, LCCRT is equivalent to cluster C, and NRT is contained within clusters A, B, and D. ORT is naturally present in the pathways designated as outliers. In Morris et al.'s study, the most common radiotherapy pattern found in the Thames Valley region was LCCRT, as in our results, and SCRT-D and SCRT-I were relatively uncommon, so it follows that these did not warrant their own cluster in this analysis. Both PORT and NRT, on the other hand, were effectively split into two groups in this analsis, which highlights that there actually exists at least two distinct subtypes within these groups

Figure 4.9 shows the characteristics of individuals in each of these major clusters. Both C and F (pre-operative radiotherapy, and post-operative radiotherapy with further surgery) has a notably higher proportion of patients being diagnosed at later stages. Cluster C also had a very high proportion of patients referred through the 2WW pathway, and a slightly higher number of patients from more deprived areas. All clusters had relatively similar age profiles, and similar deprivation levels, reflecting the relatively wealth area from which these patients were sampled. Levels of multimorbidity were also broadly similar across all clusters.

4.5 *Discussion*

The experiments described in this section demonstrate that the E-DTW measure effectively encodes both semantic and temporal information around pathways. By both objective measures (Figure 4.3) and subjective interpretation (Appendix B), the E-DTW measure's interpretation of pathway similarity matches what would be expected. Clustering using this metric identifies informative cohorts of patients that support the findings of existing research into rectal cancer treatments. In some cases, it actually provided additional insights, for example that the "no radiotherapy" group actually comprises several distinct clusters of behaviour.

This analysis has a number of limitations: the Charlson index is used here as a measure of multimorbidity, which is not strictly accurate: it only

considers conditions known to increase mortality, which is at odds with the widely accepted definition of multimorbidity which typically includes any concurrent condition, regardless of association with mortality. It is possible that patients had other conditions present that might affect treatment decisions, but which would not be affect their Charlson score.

Nevertheless, it is widely used as a general indicator of the presence of the most serious diseases (Drosdowsky & Gough 2022). In this study, Charlson scores were calculated based on the available history for each patient; at least one study has found that comorbidities can be under-reported in such data, and might therefore underestimate the Charlson score (Hua-Gen Li et al. 2019).

There was also a very high proportion of patients with unknown or unspecified tumour staging data, a long-standing issue with electronic health record data (Muller & Woods 2022). A likely solution to this problem is using NLP approaches to extract staging from the free-text radiology reports: this has already been proposed and trialled on this data (Tamm et al. 2022) but is outside the scope of this thesis.

5 *The relationship between clinical data and practice guidelines*

Clinical guidelines are typically written and disseminated as relatively unstructured text documents, so any analysis that compares guidelines to real practice needs some standard way of representing them in a structured form to accurately capture the domain knowledge they describe. There are several formal languages and structures that have been proposed for this purpose, but many are designed to support clinical decision-making in specific circumstances, rather than the representation of whole care pathways. In the previous chapter, I outlined a methodology for quantifying the difference between two patients' histories; this chapter expands on these methods and considers how to quantify the difference between a patient's history and a pathway specified in clinical guidelines.

Firstly, in Section 5.1, I examine the ways in which clinical pathways and guidelines are represented, considering their language, their structure and control flow, and their levels of completeness, and establish a set of requirements that guideline representations need to consider in the context of pathway analysis. I then discuss the existing body of work on guideline representation, focusing on computer-interpretable guideline (CIG) languages, and the BPMN and UML standards (Section 5.2), and consider how well they meet these requirements. Following this, Section 5.3 outlines a method for measuring the distance between retrospective data and guideline text. This incorporates a method for assessing the suitability of a dataset for answering a given research question, and identifying where gaps exist between the questions researchers want to ask and the data that is available; a notation for describing guideline recommendations in a structured manner; and an approach to distance measurement that builds on the E-DTW metric introduced in Chapter 4. Finally, Section 5.4 considers how this method might be applied to a specific care pathway scenario, and considers the example of rectal cancer.

5.1 *Clinical pathways and guidelines*

Conformance checking — the comparison of an idealised process described in a guideline to actual historic data — has been a long-held goal of healthcare process mining. However, as Oliart et al. (2022) observe, there are several barriers that make effective conformance checking extremely difficult in practice. Some of these are the limits of existing research: the diversity of data standards and sources make analysis very difficult, and there is so far no agreed standard method for measuring adherence. Many more, however, are actually properties of the guidelines themselves, rather than the methods or the data: a common constraint is the absence, rather than presence, of a given intervention, and they are often written in language that avoids explicit recommendations. Converting guidelines' descriptions of processes into an algorithm or a set of objective measurement criteria is therefore difficult. In order to measure the distance between the process described in a particular clinical guideline and real-world data, a guideline needs to be encoded into a computer-interpretable form.

Clinical guidelines are not the same thing as patient pathways, but as previously established in Section 3.1, they describe them and set out the many of the expected behaviours and patterns, providing a yardstick against which to measure real journeys. This section more closely examines the specific characteristics of these documents, in order to establish a set of requirements for representing them in a computable form. I consider (1) the structure and content of clinical guidelines, and the recommendations and constraints that need to be represented (2) the potential use cases of a guideline-patient distance metric and (3) the wider implications of evaluating patients' differences from guidelines.

5.1.1 *Structure and content of clinical guidelines*

As previously established (Oliart et al. 2022), the contents and layout of clinical guidelines themselves present challenges when converting them into computer-interpretable form.

Language

Boxwala et al.'s "knowledge representation framework" (2011) categorises decision support tools into four categories, which range from *unstructured* at one extreme, consisting of purely narrative text, all the way to *executable* at the other, where data elements and logical criteria are explicitly encoded in a particular syntax such that it is usable in a specific system. The vast majority of clinical practice guidelines are written as free text, and can be categorised

as unstructured according to this framework. More recently, Wyatt et al. (2023) modify this model slightly, distinguishing between fully unstructured narrative text and tagged fragments of narrative, where key words are expressly linked to specific concepts, but in practice most guidelines would remain in the first fully narrative category.

This is certainly true in the case of NICE guidelines, the primary clinical guidance available in the UK. These guidelines — introduced in detail in Section 2.3 — largely consist of individual bullet-point statements, typically consisting of largely non-committal language. A great many recommendations are non-prescriptive — the most common deontic phrase in the colorectal cancer guidelines is "consider", appearing directly eleven times. Where recommendations are more prescriptive, they are still not compulsory — "must" never appears at all, but the second most common phrasing was "offer", appearing nine times, which indicates a stronger instruction but still implies a patient's right to decline. The particular choice of language[9] is intended to reflect the strength of evidence (Garbi 2021).

It is therefore often challenging to measure how well some of these recommendations are followed: a patient record will rarely reflect whether a treatment option was "considered by the clinician", and there is also relatively formal decision criteria or logic in the guidelines, although where there is this usually takes the form of a reference to an external guidance document.

9 How clinicians interpret this language is a different question: separate research suggests that "must" is commonly read to mean the strongest possible obligation, and "may" the weakest, but the meaning of anything in between (e.g. "should", "recommended", "suggested") is interpreted inconsistently (Lomotan et al. 2010).

Structure and control flow

One way in which guidelines are distinct from pathways is that in many cases, a treatment pathway can often be described by multiple guidelines. For example, most cancer treatment begins with the steps described in the guideline *Suspected cancer: recognition and referral* (NICE 2015). This describes the steps used in primary care settings to identify all types of cancer, but is not an exhaustive description — for example, the guidelines for FIT, an essential part of the modern colorectal cancer pathway, are detailed in a separate guideline (NICE 2023c). Post-referral steps are described in guidelines specific to each cancer site, such as *Colorectal cancer* (NICE 2020), where readers are again referred to a separate document for advice on specific treatments on at least four occasions. Therefore, a patient's entire journey from the start of symptoms to final outcome is not described in one place, and is distributed across several different guideline documents.

Furthermore, the statements in guidelines are not always organised in chronological order or grouped by clear treatment state; since the retirement of the NICE Pathways service in 2021, there is no official flowchart representation of these guidelines. Some include flowcharts, but these are not

necessarily formal or useful models of the treatment pathway: Scott et al. (2023) describe the flowcharts used in one set of NICE guidelines as "more of a psychological model than a strictly logical model … depicting clinical thinking and problem-solving rather than a fully specified operational process flow".

Localised pathways disseminated by individual organisations will generally have a more deterministic flow, with specific decision criteria and recommended treatments. For example, the Thames Valley Cancer Alliance's colorectal cancer guidelines lay out the possible routes in a flowchart (TVCA 2021, 2023). These local pathways are however written by individual organisations, and therefore do not follow any standard notation or format, making comparison difficult.

Guidelines that take a more holistic view of the whole pathway do exist; in particular, national pathway guidelines tend to place great emphasis on timing constraints, especially in cancer care, where providers are often evaluated based on how quickly they handle patients with suspected cancer. Historically, this has meant the two-week wait (2WW) standard, which recommends that the time from a patient's referral by a GP or screening programme to their first specialist appointment should be no more than fourteen days, but this is increasingly being replaced by the *faster diagnosis standard*, which states that a diagnosis should be either made or ruled out within 28 days (NHS England 2023a). For this reason, most national documentation that describes the cancer pathway focuses on the timings of these investigative and diagnostic stages, and end at the point of first treatment (NHS England 2023b; NHS Wales 2023). The Aarhus model, an international standard for describing the key time points and delays in the cancer pathway, similarly focuses on the diagnostic stages and ends at treatment (Weller et al. 2012). M. Morris et al. (2020) however build a fuller conceptual model of the cancer pathway, and divide it into four main stages: pre-diagnostic, diagnostic, treatment/management, and survivorship.

Incompleteness

Guidelines tend to cover a small number of high-impact scenarios, rather than exhaustively covering an entire clinical topic (Garbi 2021). This distinguishes them from pathways in the sense that a guideline does not cover all possible treatments and steps — just the most common or the most variable ones.

Even where many separate guidelines cover a particular area, there are still gaps: in the NICE colorectal guidelines previously discussed, the referral guidelines end with a patient's referral to secondary care, and the colorectal cancer treatment guidelines begins divided into sections depending on the

cancer site, which implies the presence of an intermediate investigation and diagnosis phase, but does not describe or detail it. By contrast, many of the more localised guidelines such as the TVCA guidelines do cover this phase.

The TVCA, NHS England, NHS Wales, and Aarhus models all incorporate a "treatment" stage without any particular description of what this comprises. These models presumably avoid describing the actual treatment part of the pathway because it is incredibly heterogeneous, and specific to individual patients. The three most common components of treatment for colorectal cancer are surgery, chemotherapy, and radiotherapy, but the specific combination and order of these three components will vary. Each can also vary in different ways: the particular type of surgery used; the combination of drugs, number of cycles, and timing in chemotherapy; and the level of radiation, number of fractions, and timing of radiotherapy. The investigative stages, by comparison, are relatively standardised, and often contain large portions that are common to multiple cancer sites, so much greater attention has been paid to mapping them. As well as the obvious factors, such as the patient's stage of disease and the type of cancer, there are also a myriad of contextual factors that affect treatment: some of these, such as treatment preferences, are within the patient's control, some, such as age, gender, ethnicity, or deprivation, are not, as demonstrated in Chapter 3.

The treatment process is not decided by a single person or algorithm. In the UK, the accepted standard for managing a cancer patient's treatment is the multidisciplinary team (MDT), a regular meeting of a number of different clinicians including oncologists, surgeons, radiologists, nurses, and histopathologists, who examine each patient's history and comorbidities, imaging and histopathology reports, and psychosocial factors and treatment preferences (Soukup et al. 2020). Precision medicine — the idea that treatment should be very closely tailored to a patient, often including predicting the best possible option on the genetic level — is an increasingly popular research area. Based on this, it appears that the treatment stage will only become more personalised in the future (Jameson & Longo 2015).

For these reasons, analyses of treatment patterns usually focus on a specific sub-population, such as rectal cancer or a specific stage colon cancer, and generalise treatments into broad groups. E. J. A. Morris et al. (2016), for example, consider only rectal cancer patients who received surgery, and categorise the most common patterns of category and surgery into groups. This makes it very hard to compare a specified treatment pathway against the real-world data when the treatment stage is heterogeneous and inconsistently defined.

5.1.2 *Use cases*

A summary of treatment processes, of the sort typically avoided by guidelines, is useful for a number of reasons. As discussed in Chapter 4, the exact combination and ordering of surgery and radiotherapy for rectal cancer is the subject of debate; examining the patterns of usage, the natural clusters of patterns, and the relationship between these patterns and patient characteristics provides useful insight into clinical practice. In colon cancer, there are similar questions around the usage of adjuvant chemotherapy — guidelines vary significantly on the recommended drugs and duration of treatment, and different studies have suggested that effectiveness varies, or doesn't, by age. Because of this, individual teams in NHS trusts are responsible for making these decisions in England, and those decisions can vary significantly (Boyle et al. 2020; Taylor et al. 2021). Recommended pathways also change over time: in particular, the COVID-19 pandemic led to several temporary modifications to diagnostic and treatment pathways (E. J. A. Morris et al. 2021). Among the recommendations made at the time were that a large number of non-essential or emergency colonoscopies should be deferred for patients on the 2WW pathway, and that new diagnoses should not be treated unless complications require emergency admission (Fearnhead et al. 2020); that patients with lesser urgency and higher risk from COVID should be deferred (BASO 2020); and that short-course radiotherapy should be preferred over other radiotherapy options such as chemoradiotherapy, in order to reduce appointments and contact with staff and avoid the negative effects of chemotherapy on immune system function (Marijnen et al. 2020).

Whilst process mining methods have been applied to healthcare processes at varying levels of granularity, these research questions suggest that the major interventions that make up a journey should be the focus of analysis, rather than the more specific events such as individual measurements and ward movements which are more often the focus of PM research.

5.1.3 *Do we want conformance checking in healthcare?*

Patient pathways and treatment guidelines are not without their issues. Writing a single procedure for a complex disease inevitably requires simplification; in particular, most guidelines are themed around one particular condition, ignoring potential conflicts between guidelines in cases of multimorbidity (Dwyer et al. 2023). The concept of a single pathway for each disease sits at odds with the fact that individual patients have different priorities and wishes, and value sharing in the medical decision-making process; clinicians are expected to offer different treatment based on the case in front of them

and their professional judgement. Pathways represent recommended routes of care, but are explicitly not compulsory routes of care, and any analysis of patient pathways therefore needs to consider them in this context.

The economist Charles Goodhart is credited with coining "Goodhart's law": the observation that "when a measure becomes a target, it ceases to become a good measure" (Lamba 2021; Mattson et al. 2021). Hospitals are often, for example, encouraged to reduce patients' length of stay where possible, but a single number does not tell us whether a patient was discharged when they *should* have been given their individual circumstances. Encouraging and incentivising healthcare providers to meet particular targets may result in them modifying their behaviour in ways that meet the target, but miss the point. Pathway "compliance" or "comformance" is therefore neither realistic nor desirable to enforce. This is relevant when we are attempting to develop an algorithm that quantifies the distance between a patient and a guideline. Analysing the pathways of patients is — as the previous chapters have proven — undoubtedly a useful endeavour that enables insights to be drawn into factors that affect patient treatment and outcomes, and whether or not a particular journey resembles the expected one is another facet of this. Clinical pathways provide a useful analysis framework because they describe the *expected* or *typical* treatment for a particular disease. It is important then to emphasise that when this chapter makes reference to "distance" or "divergence" from an ideal pathway, this does not carry a value judgement, or imply an error to be corrected; it is the starting point of a further investigation into factors that affect clinician decision-making and ultimately patient outcomes.

5.1.4 *Requirements*

Whilst the individual recommendations within NICE guidelines are considered gold-standard treatment recommendations, the absence of decision criteria or clear patient flow, and entirely missing stages means that there is a high degree of domain knowledge and professional judgement assumed on the part of the clinician, and they do not form full "pathways" in the sense of our definition. Since the recommendations described in clinical guidelines are more often than not an unstructured set of recommendations without order, arranging theme into a flowchart of the expected pathway requires knowledge of the specifics of a particular disease. It can reasonably be assumed that there exists some sort of investigative or pre-diagnosis stage, followed by the diagnosis itself, followed by the main treatment, and then further or follow-up treatment, but the ordering of events beyond this is de-

pendent on the particular disease. Relatively few recommendations describe the entire pathway in this way.

It has also been established that binary notions of pathway or non-pathway behaviour are too rigid, and that process "conformance" needs to be measured across multiple dimensions, including the ordering of events, the timings between them, and the similarities between them. I have already outlined the E-DTW method (Chapter 4), which can compare pathways according to these factors; this section therefore considers how to encode guideline pathways into a compatible form.

From examining the language, structure and control flow, and incompleteness in clinical guidelines (Section 5.1.1), it is apparent that guidelines commonly reference other guidelines and documents, contain missing or assumed segments, and are not always arranged in chronological order, meaning that a language must support the composition of pathway models from multiple guidelines. Considering the use cases for pathway analysis (Section 5.1.2), it is also important that a language support the description of events at the correct level of granularity for pathway use cases, i.e. major procedures rather than granular or minor events. It is also clear, through examination of both the text of clinical guidelines and their related research questions, that timing is highly emphasised in both. Any language therefore needs to support the description of timing constraints between events. In addition, it is important to support transformation into a form compatible with the E-DTW method outlined in Chapter 4, so that "compliance" with a pathway can be measured as a continuous distance rather than a binary variable.

5.2 *Existing approaches*

Several previous approaches have been proposed for encoding information around medical processes; these can be summarised according to two main groups.

5.2.1 *Representing pathways with computer-interpretable guidelines*

Computer-interpretable guidelines (CIGs) systems are a family of methods for representing clinical practice guidelines in forms that can be processed and reasoned over by a computer. CIG languages take a range of different forms, but the logic they rely on can be characterised either as *semi-formal model*, which describes a guideline without executing it, or a *formal model*, where decision criteria are represented by logical constructs and bound to

concrete patient data values, and can generate decision outcomes in the form of recommendations (Peleg 2013).

One of the earliest CIG languages was the Arden Syntax, introduced in 1989 (Oliveira et al. 2014). Here, medical knowledge is encoded as "medical logic modules", with each individual module containing enough rules to make a single decision. This relatively simple format limits scalability, and makes it difficult to properly represent an entire guideline, or by extension a clinical pathway. Guideline Interchange Format (GLIF), introduced in the 1990s, represents a sequence of steps in a flowchart style, with each step either a medical action, an activity-oriented action (messaging, retrieving data), or a control action (invoking sub-guidelines). Additionally, steps representing patient state describe an individual's health condition and function as data entry points, so that once the state of a patient is updated, the guideline executes accordingly.

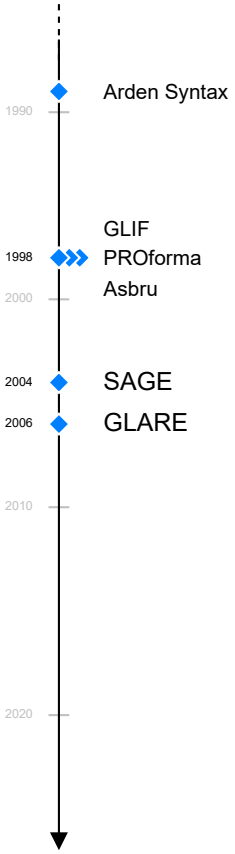

Figure 5.1. Timeline of major CIG languages

PROforma (1998) similarly represents guidelines as flowcharts, with each node an instance of a pre-defined task class, but every task derives from a common "root task", and a plan can contain any number of atomic tasks, meaning that a single guideline is effectively a hierarchy of nested tasks. Notably, PROforma was the first formalism to support uncertainty in the decision-making process. A decision object associates positive and negative signs with each logical expression, which gives each argument its weight in the candidate's overall score. Asbru (1998) describes guidelines in a similar way, defining its "plans" as a hierarchy of tasks. Clinician actions can also be assigned temporal constraints known as "intentions", which can represent a state to be maintained, reached, or avoided; an intermediate action which must be performed during the plan; a patient state that must be verified at the end of plan execution; or the pattern of clinician actions that should result from the plan. The intended or expected effects of the plan can also be expressed, which can be used to describe how the plan arguments and measurable parameters affect each other. Asbru also heavily emphasises the temporal aspects of pathways, with each plan having four different time properties: earliest and latest starting shift, earliest and latest finishing shift, along with minimum and maximum durations.

GLARE (Guideline Acquisition, Representation and Execution) describes a graph-like structure for CPGs, with clinical actions represented by nodes and decision criteria represented as (diagnosis, parameter, score) triples. Atomic actions representing simple tasks such as queries, actions, and decisions can be combined into composite actions. SAGE (Standards-based Shareable Active Guideline Environment), is similar to GLARE, representing guidelines as graphs of nodes, but emphasises supporting guideline sharing across het-

erogeneous information systems, meaning interoperability of clinical data is supported through the HL7 and the UMLS standards.

Mulyar et al. (2007) evaluate the capability of CIGs to represent *workflow patterns*, a set of common control-flow patterns identified in the business process management literature. PROforma offered the greatest support (22 patterns out of 43), whilst, Asbru, GLIF and EON supported fewer. This is not necessarily a useful measure of CIGs' effectiveness: Kaiser & Marcos (2016) argue that not all workflow patterns are needed to describe guideline knowledge.

Many of these CIG languages have drawbacks that have prevented their widespread adoption outside of academia (Greenes et al. 2018; Oliveira et al. 2014). In particular, most fail to successfully strike a balance between complexity and expressive power: a simple model may not be able to represent all of the information in a guideline — as in Arden's single-decision modules — but it can also be too complex to be practically useful — as in PROforma and its many proprietary specifications for data. Greenes et al. (2018) argue that there are fundamentally so many different perspectives involved in clinical guidelines that designing a single language is a futile pursuit. It is therefore important to specifically evaluate their suitability for answering pathway questions.

CIG languages are normally designed to be used in conjunction with a *guideline execution engine*, which accepts data input at the point of care and produces a treatment recommendation for a clinician, and these languages are therefore designed and evaluated with that task in mind (Isern & Moreno 2008). Relatively little literature considers the other practical applications of modelling guidelines, for example being able to compare guidelines against real patient traces on a population scale. Van de Klundert et al. (2010) develop a set of algorithms for measuring pathway adherence, but note that executing this method over thousands of patient instances would take time in the order of weeks. Peleg (2013) discusses such a use case, but describes it in terms of comparing EHR data to CIG execution logs, to evaluate whether clinicians followed the recommendations given by their decision support systems. This sort of evaluation would require every clinician in the pathway to use a CIG for every decision of interest, and every clinician to use the same CIG system. Additionally, introducing or enforcing such a system for study purposes could potentially cause changes in clinical practice; the process of data collection would actually interfere with the process being studied, undermining its validity.

Given the age of many of these languages, very few CIG toolchains are still publicly available online, and fewer still are free and open-source (Gamba

2017). Therefore, whilst computer-interpretable guidelines are useful tools at the point of care, they are largely not suitable for comparison against retrospective EHR data, which would be needed to enable large-scale studies of pathways.

5.2.2 *Representing pathways with process languages*

Given that CIG approaches tend to focus on producing decision aids for individual interactions with the healthcare system, researchers aiming to model entire pathways have also made use of pre-existing, domain-agnostic languages for describing processes. The two most common are BPMN and UML.

BPMN

Business Process Model and Notation (BPMN) is a graphical model for specifying business processes. Its primary stated aim is to create models that are readily understandable by all users, and to provide a standardised bridge between business process design and implementation (OMG 2013). BPMN, as a widely-used standard for communicating processes in the business world, is therefore a natural candidate language for communicating clinical pathways.

BPMN models are composed of several core elements. *Events* and *activities* represent events that occur and actions that are performed by the business, whilst *sequence flow* arrows connect the elements in order of execution, and *gateways* control the divergence and convergence of sequence flows. As well as these core elements, there are a wide variety of additional elements specified in the BPMN standard, such as *pools* and *lanes*, which partition elements according to an individual or organisation's separate responsibilities, variants of the gateway for more complex control flow (*exclusive, inclusive, event-based, complex* or *parallel*), and additional arrows that represent message passing between participants, or link additional artifacts and annotations to elements.

Extensions have been proposed to adapt BPMN to the needs of healthcare processes. Chae et al. (2020) convert pathways to BPMN by classifying pathway elements into categories, then mapping each concept to its nearest equivalent in BPMN. Braun et al. (2014) follow a similar process, categorising pathway concepts by how similar they are to existing BPMN elements, but also propose a number of extensions including specifying the Task concept into diagnosis, therapy, and supporting tasks, and an evidence-based gateway to annotate decision criterion and logic.

However, there are also drawbacks to BPMN. The specification is at times ambiguous, leaving space for process interpretations that theoretically conform to the standard but are incompatible in practice. For example, where multiple interrupt events are nested, there is no rule as to whether one or all should execute, and in which order. Concepts not represented in BPMN include state and structure, as well as any mechanism for refining conceptual models to executable ones. Another key issue is the very large number of constructs that BPMN defines, many with unclear overlaps in functionality, and able to be defined in terms of each other (Börger 2012). The language is incredibly complex, consisting of a large number of constructs, relatively few of which are regularly used in day-to-day practice (zur Muehlen & Recker 2008). BPMN's graphic notation often creates overwhelming and confusing diagrams which hamper communication with stakeholders and domain experts (Genon et al. 2010).

These issues are not insurmountable: Scheuerlein et al. (2012) describe a pilot project in which clinicians modeled processes with BPMN, and Kirchner et al. (2023) describe the development of an intermediate language for subject matter experts to develop pathways in before converting them into BPMN. However, the fact that the former involved a one-day introduction to BPMN and that the latter required an entire intermediary language to be created do not dispel the idea that BPMN is complex. This complexity is two-way: Pufahl et al. (2022) attribute the lack of adoption of BPMN in healthcare to the massive complexity of the healthcare processes themselves, along with the relatively slow adoption of technological innovations in healthcare, which itself further complicates the processes.

Most importantly for our purposes, BPMN offers limited support for temporal modelling. The constraints used in healthcare process are more complicated than BPMN can express (Pufahl et al. 2022), and whilst extensions have been proposed for handling temporal features (Cheikhrouhou et al. 2013; Gagne & Trudel 2009), no one has become universally adopted or accepted.

UML

Unified Modelling Language (UML) is a general-purpose modelling language, predominantly used for analysing, describing, and designing systems and software (OMG 2017). In terms of control flow and data perspectives, UML is capable of describing business processes to the same extent as BPMN, with most of the missing features concerning resource or organisational aspects (N. Russell et al. 2006; Wohed et al. 2006). Much like BPMN, UML has been

applied to healthcare contexts (Gencturk et al. 2024; Liyanage et al. 2016; Luzi et al. 2019; Pecoraro & Luzi 2022).

UML incorporates two notations that can encode processes (Rumbaugh et al. 1999). The most widely known notation, and the one used by most of the aforementioned sources, is the *activity diagrams*, which resembles a conventional flowchart, consisting of different types of nodes (actions, decisions, and splitting or joining of concurrent activities), linked by arrows, together showing the activities required to perform a calculation or process. These activities can be sequential or concurrent, and can have input and output parameters that are passed to each other. However, the UML standard also supports *state machines*, in which nodes represent states, and edges denote transitions. The state machine can be seen as a stricter subset of the activity graph, modelling all possible life histories of an object. Unlike an activity, a state's behaviour depends only on its current state, without carrying any variables, but the triggering of a transition can be predicated on a particular condition.

Evaluation

Most of these approaches that adapt or extend BPMN and UML for clinical guidelines stop at representation, and do not consider conformance checking as an application. Savino et al. (2023) do however compare guidelines to data using a different language, Pseudo-Workflow (PWF). PWF's authors do not provide a detailed semantics (Gatta et al. 2017a,b), but it is relatively simple, consisting only of nodes, edges, and edge conditions, making it similar to UML's state diagrams. Savino et al.'s analysis is striking in that it observes relatively few patients following the exact sequence of events outlined a guideline. This highlights the limitations of a purely logic-based conformance checking approach — almost every patient will deviate from a rigid flowchart in some way or another. Incorporating "soft" constraints, or describing different levels of deviation would lead to more informative analysis.

These existing approaches do not yet address the requirements of patient pathway analysis. Many focus on the description of events at a very granular level and notations such as UML and BPMN have limited or inconsistent support for timing constraints. Fundamentally, logical approaches that evaluate conformance are fundamentally rigid in their approach. For this reason, this chapter investigates the feasibility of an alternative approach: measuring guideline-pathway similarity as a continuous distance.

5.3 *Methods*

This section describes a framework for representing and comparing recommended treatment pathways against retrospective EHR data. As the previous section has established, clinical process data is inherently heterogeneous and unpredictable; strictly logical approaches to conformance checking tend to naturally highlight that very few patients exactly follow the pathway. This method therefore adapts the E-DTW approach described in Chapter 4, with the aim of achieving a metric that can describe patient-guideline similarity as a continuous score, usable as input for a range of different analyses. Given a particular recommended or guideline pathway, and a retrospective EHR dataset, it returns for each patient a score measuring the similarity of each patient to the model pathway, allowing a user to measure the extent to which real-world pathways resemble guidelines.

The method consists of three steps: firstly, an assessment of a retrospective dataset, focusing on the extent to which it actually records the events mentioned in clinical pathways; secondly, patient pathways are recorded in a purpose-built notation; and finally, these encoded pathways are transformed into a computable form to which E-DTW can be applied.

5.3.1 *Assessing retrospective datasets*

Before any attempt to analyse the relationship between ideal care pathways and real data, it is first important to understand how suitable the available dataset is for answering the research question of interest. Martin et al. (2024) consider an important question regarding EHR data: what actually gets recorded? Their study observes 38 distinct activities taking place, only four of which actually leave an explicit trail in the EHR. This study — whilst focusing on a much more granular type of event than in this thesis' definition of patient pathways — establishes an important principle: there exists a mismatch between events that happen in practice, and events that get recorded. This is, to some extent, inevitable: clinicians carry out many thousands of activities every day, so it is natural that some will be prioritised for recording. However, it does inspire a further question: does there also exist a mismatch between what is expected or recommended in guidelines, and what is recorded? In other words: how many of the recommendations made by guidelines are actually visible in the data?

In this method, each recommendation in a guideline is assigned a *visibility* value (*true* or *false*) based on whether or not the core event in that particular recommendation is visible in our dataset. Where visibility was *true*, recommendations were further rated each according to the ease with

which the event it concerns can be extracted from the dataset into an event log format:

A The event is directly visible as a single event in the patient record, for example "surgery" or "colonoscopy"

B The event is visible with a small amount of abstraction or transformation; for example, it is indicated by a sequence of multiple single events, or two subsequent events separated by a particular time window

C The event requires heavy transformation or the application of logical rules, for example reference to the occurrence or ordering of other events, or looking up test results

Furthermore, the presence of the event's preconditions or reasoning was also recorded as a binary property. For example, the recommendation "All patients with suspected large bowel obstruction should have a contrast-enhanced CT" (Moran et al. 2017) received a score of "A" for ease of event extraction, since a patient clearly either does or does not have the OPCS code for CT scan in their procedure history, but its reasoning was marked as *false* since the clinician's suspected diagnosis, if there was one, is not clearly visible within the structured data available.

Comparing the clinical guidelines against the actual data available ensures that the method focuses on events at the right level of granularity, and that the method is able to produce results that are as meaningful and close to complete as possible.

5.3.2 *A lightweight notation for patient pathways*

Once it is established that all or part of a pathway is visible in a dataset, the pathway needs to be converted from its free-text or non-standardised form into something computer readable. Here, I propose a notation for describing patient pathways, which is based on the UML state machine standard, rather than BPMN, for a number of reasons:

- Firstly, *homogeneity*: the ability to represent a process very simply, as a set of states and transitions between them with few extras makes models easier to create, easier to interpret, and easier to convert into alternative representations such as graphs. This notation follows the process mining convention of referring to a retrospective dataset as an *event log*, fundamentally composed of only one basic object type: the event.
- Secondly, *temporality*: both BPMN and UML have limited official support for temporal constraints, which has been an issue in previous attempts at codifying medical guidelines. Several extensions to both languages

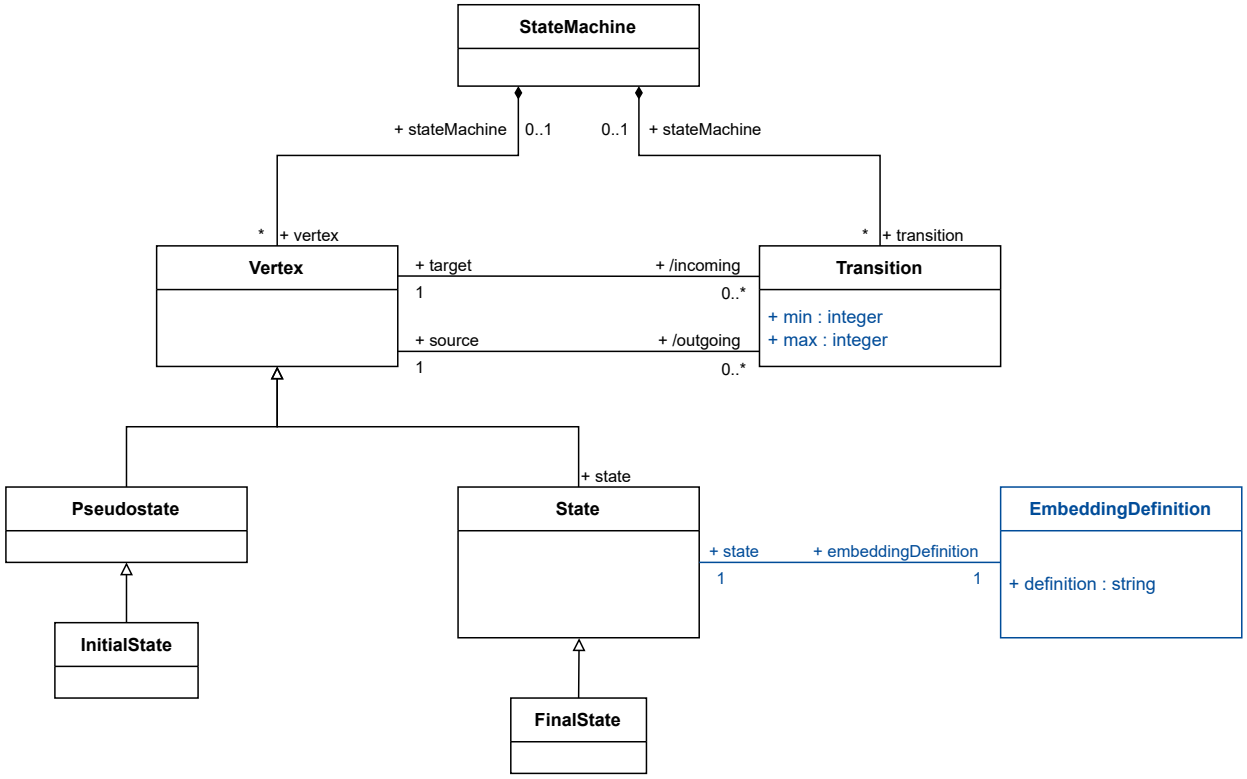

Figure 5.2. The proposed patient pathway notation, based on a subset of UML state machines

introduce a wide variety of timing constructs, but in practice the most significant constraint involved in patient pathways is the time taken to move between events.

- Thirdly, the *atomicity* of transactions: a transition between two events — although it may take a period of time — should be a single, atomic, action without asynchronous communication or message passing: from the point of view of a retrospective dataset, an event either follows another event, or does not.
- Finally, the handling of *choice*: again, in a retrospective dataset, an event is either followed by another event, or it isn't: decoration such as parallel or logical gates is not necessary.

Other process languages exist as potential bases for this language, but all come with both advantages and disadvantages. Petri nets were discounted because of the lack of concurrency as a requirement. There are also many languages proposed in the field of declarative process mining, but no single widespread standard (Hanser et al. 2016; López & Simon 2022).

The basic UML state machine standard (OMG 2017) is detailed in Section 2.6. Many of the more complex state chart concepts tend towards describing an executable or programmatic implementation of a particular process, whereas the focus of this research is on the recording of historic events. It is

therefore possible to significantly simplify some of these elements. In order to further tailor this notation the needs of pathway research, this section defines a limited subset of this standard, and introduces several useful constraints. The simplified pathway state machine notation is summarised in Figure 5.2.

UML *states*, rather than *events*, are used to represent clinical events from the EHR, since clinical events found in event logs more closely resemble states in that they are not instantaneous and have extended durations (for example, chemo- or radiotherapy). Only simple states are used, as analogous to the simple atomic events found in event logs. The notation of a *pseudostate* — a state-like vertex with different execution semantics — is retained, since whilst it has no clear analogue in patient pathways itself, it forms the basis for the definition of the *initial state*, which is definitely useful.

Even if it is possible in theory for two activities to occur in parallel, it is not a requirement that these concurrent activities be modelled. This is a well-established principle used in older process languages such as CSP. As Hoare (1985) puts it, defining the notion of a process's *trace*:

> Imagine there is an observer with a notebook who watches the process and writes down the name of each event as it occurs. We can validly ignore the possibility that two events occur simultaneously; for if they did, the observer would still have to record one of them first and then the other, and the order in which he records them would not matter.

In this use case, where comparison is made against historic EHR data, this is literally the case. The trace is all there is to work with: if two events, A and B, could in theory execute in parallel, they will in practice appear in data as either (A, B) or (B, A). These possible transitions are therefore included in the UML model when it is created.

In the original specification, transitions have an undefined duration, allowing them to be either instantaneous or take time. In this interpretation, transitions may — and often do — have non-zero durations, since they represent the "waiting time" between clinical events, which is of great interest for research. For this reason, the definition of a transition is extended to add two attributes: a minimum and maximum acceptable duration. Triggers and constraints are removed from transitions, since the preconditions for various events and the reasoning used by clinicians are rarely visible in the data, as will be discussed in the following section.

All transitions are assumed to be *completion transitions*, i.e. transitions that are automatically entered once the source state has completed; once

a patient undergoes an event, they are considered to have automatically entered the "waiting time" for the next event. Two attributes are added to each transition: a minimum and a a maximum acceptable duration for that transition, i.e. the time spent waiting between the two events. One extra entity is introduced: the *embedding definition*, which maps a particular state to a text description or a section of pseudocode or code that describes the process of generating embeddings compatible with the E-DTW metric.

The later stages of the proposed method involve generating embedded representations of possible pathways, which entails generating graph traversals over the state machine to get all possible valid paths. This constraint therefore removes the risk of creating infinite loops. If an event can validly repeat itself — for example, cycles of chemotherapy — this should instead be represented as several possible embedding mappings, representing the set of reasonable possibilities.

The resulting notation is founded on the UML standard, allowing for a consistent definition, but tailors it towards the needs of analysing pathways in EHR data by eliminating elements that tend towards representing executable or programming implementations of processes, and introducing attributes that support compatibility with the E-DTW metric.

5.3.3 *Comparing pathways and guidelines in embedding space*

Given a dataset that records events of interest for the pathway, and given an encoded, computer-readable pathway, the final step is to measure the distance between the two. In this proposed method, a guideline is converted into a set of traces representing exemplar patients that comply with the guideline, which are subsequently converted into a low-dimensional representation. Real patient pathways are ranked based on their distance to their nearest exemplar patient, according to the E-DTW measure.

Generating patients

The guideline, encoded as a UML state machine, is treated as a directed graph G, where activities are nodes and transitions are edges. In practice, this can be achieved by encoding the UML representation in a text-based syntax (PlantUML 2023), in which transitions are simply written `A -> B`; these can be easily converted to $(A, \text{followedBy}, B)$ triples. Transitions can optionally be annotated with a tuple $(\max, \min)$, which denotes the maximum and minimum acceptable time delay between the two events. A "valid" trace through the guideline g_i is defined as any simple path across the graph G from the *start* to the *end* node — that is, a sequence of adjacent edges such

that no nodes are repeated. Paths can then be converted into one or more exemplar timeseries $I(g_i)$: these are created by reading each node and edge from start to end, and replacing each node with a vector or sequence of vectors representing the event.

Each edge is treated as a run of zeroes: if a transition has been labelled with a $(\max, \min)$ tuple, the run's length is randomly sampled from this range; if none is specified, it is sampled from a default range of $(0, 28)$.

Comparing against patient traces

For each unique path through the guideline g_i, 20 exemplar timeseries $I(g_i)$ are sampled. A distance matrix is calculated between every exemplar and every real patient in the dataset, with distance being defined using the E-DTW measure outlined in Chapter 4. In these experiments, the E-DTW $(nd, 0)$ configuration was used, based on several factors. nd approaches typically better matched semantically similar events than the $2d$, and the penalty $(+p)$ was omitted to allow for a more flexible matching of time, given that in many cases there was a range of acceptable possible time values. Given the relatively small differences between 0 and i variants, the less computationally intensive option was chosen.

For each patient p_j, their *minimum guideline distance* (MGD) is defined as the shortest E-DTW distance between that patient's pathway and any of the guideline instances:

$$\text{MGD}\left(p_j\right) \;=\; \min\left(\text{dist}\left(p_j, I\left(g_i\right)\right) \;\forall\; g_i \in G\right)$$

The cohort of patients was the same as described in Section 4.3.1; for each patient, their pathway consisted of any events defined as *major resection*, *minor resection*, or *diagnostic testing/imaging* according to the extended CORECT-R codelist, since these were the possible events described in the guidelines under examination.

5.4 *Application*

This section examines actual treatment guidelines for colorectal cancer, and uses the proposed methodology to evaluate them and map them into a UML-based notation.

5.4.1 *Data assessment*

As in previous chapters, the data available for research was based on the NIHR Health Informatics Collaborative colorectal cancer project (Tamm et al. 2022). The data was compared to three sets of clinical guidelines.

The National Institute for Health and Care Excellence (NICE) guidelines on colorectal cancer (NICE 2020) consisted of 39 bullet-pointed recommendations, of which 15 (28.3%) were visible in the data. Of these 15, 9 (60.0%) were rated "A" for ease of extraction, and 5 (33.3%) involved a set of preconditions that were not measurable or recorded in the dataset.

The Association of Coloproctology of Great Britain and Ireland (ACPGBI) guidelines (Geh et al. 2017) contained 135 distinct recommendations in the *diagnosis, investigations and screening*, *surgical management*, *multidisciplinary management*, and *follow up, lifestyle and survivorship* sections. 63 (46.0%) were visible, and of these 26 (41.27%) were rated "A", and 8 (12.7%) involved a precondition not visible in the data.

The European Society for Medical Oncology (ESMO) guidelines on localised colon cancer and on rectal cancer (Argilés et al. 2020; Glynne-Jones et al. 2017) contained 53 individual recommendations, of which 32 (60.0%) were found to be visible in the data. Of these, 22 (68.8%) were rated "A", and 7 (21.9%) involved a set of preconditions that were not measurable from the dataset,

5.4.2 *Notation*

The ESMO guidelines were selected for application, since they had the largest proportion of recommendations visible in the dataset. They also resolve some of the guideline shortcomings previously identified: most notably, they describe the options available at the treatment stage in detail, and provide these recommendations as near-complete flowcharts. The treatment sections were therefore chosen as they were the most complete specifications of recommendations in any of the guidelines examined. Focusing on these major treatment decisions ensures that the method is examining the events of most importance to treatment, and therefore of interest to research, and focusing on the correct level of granularity.

Each of the major recommendations — i.e. those contained in the "rectal cancer treatment" flowcharts — was summarised as a simple statement ("event A should be followed by event B"), then encoded as in the model as a transition from a state representing event A to a state representing event B. By breaking guidelines down into their component recommendations,

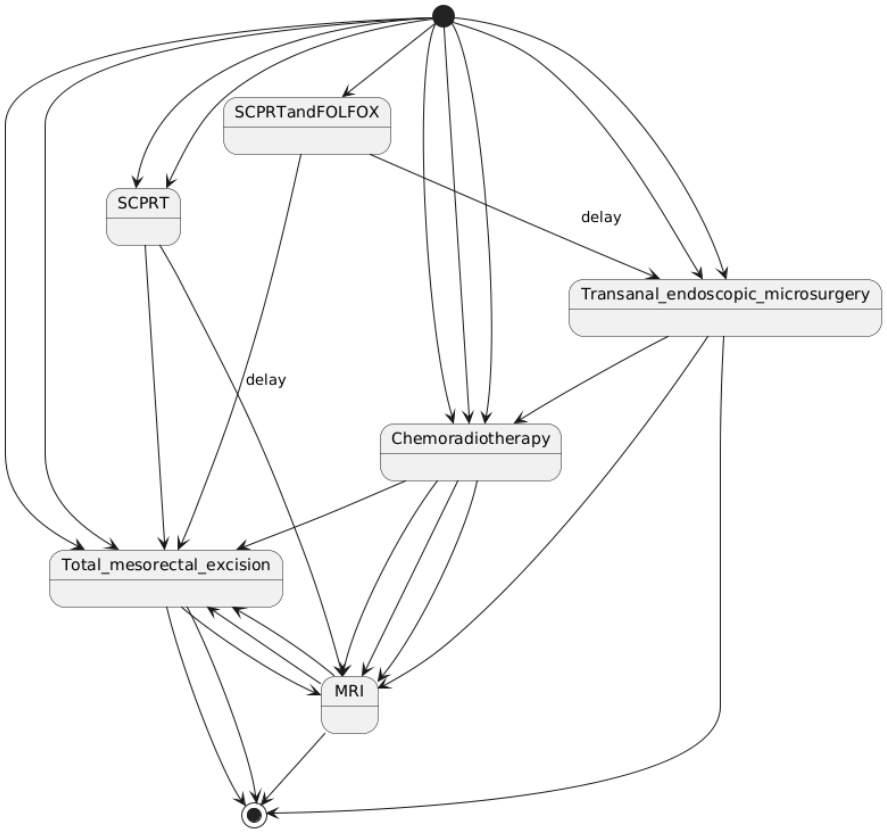

Figure 5.3. UML model of possible rectal cancer treatment pathways, according to the ESMO guidelines

a more complete model can be composed from statements across different sections and guidelines. The complete model is shown in Figure 5.3.

5.4.3 *Comparison*

Converting guidelines from the UML-like notation to a timeseries form required some further interpretation, in order to identify appropriate representations for clinical events. Simple events, i.e. those equivalent to a single concept and occurring at a particular instance in time (i.e. anything less than one day in duration, since this is the level of quantisation of the embedding timeseries), for example imaging or surgery, were encoded by a single embedding representing the relevant SNOMED concept.

More complex events were defined based on domain knowledge. For example, the ESMO guidelines refer to short-course preoperative radiotherapy (SCPRT); all of the sources cited in support of this recommendation defined SCPRT in the same way, as 25Gy of radiation delivered in five fractions of 5Gy each[10], so SCPRT was defined represented as a five-day block of the *radiotherapy* SNOMED embedding. Chemoradiotherapy was more complex since the sources cited by the guidelines did not define it consistently: whilst 45-50Gy of radiotherapy was typically used as a base, it sometimes entailed chemotherapy in the first and last weeks, sometimes chemotherapy on

10 One *gray* (Gy) is a unit of ionising radiation equal to one joule of energy per kilogram of matter.

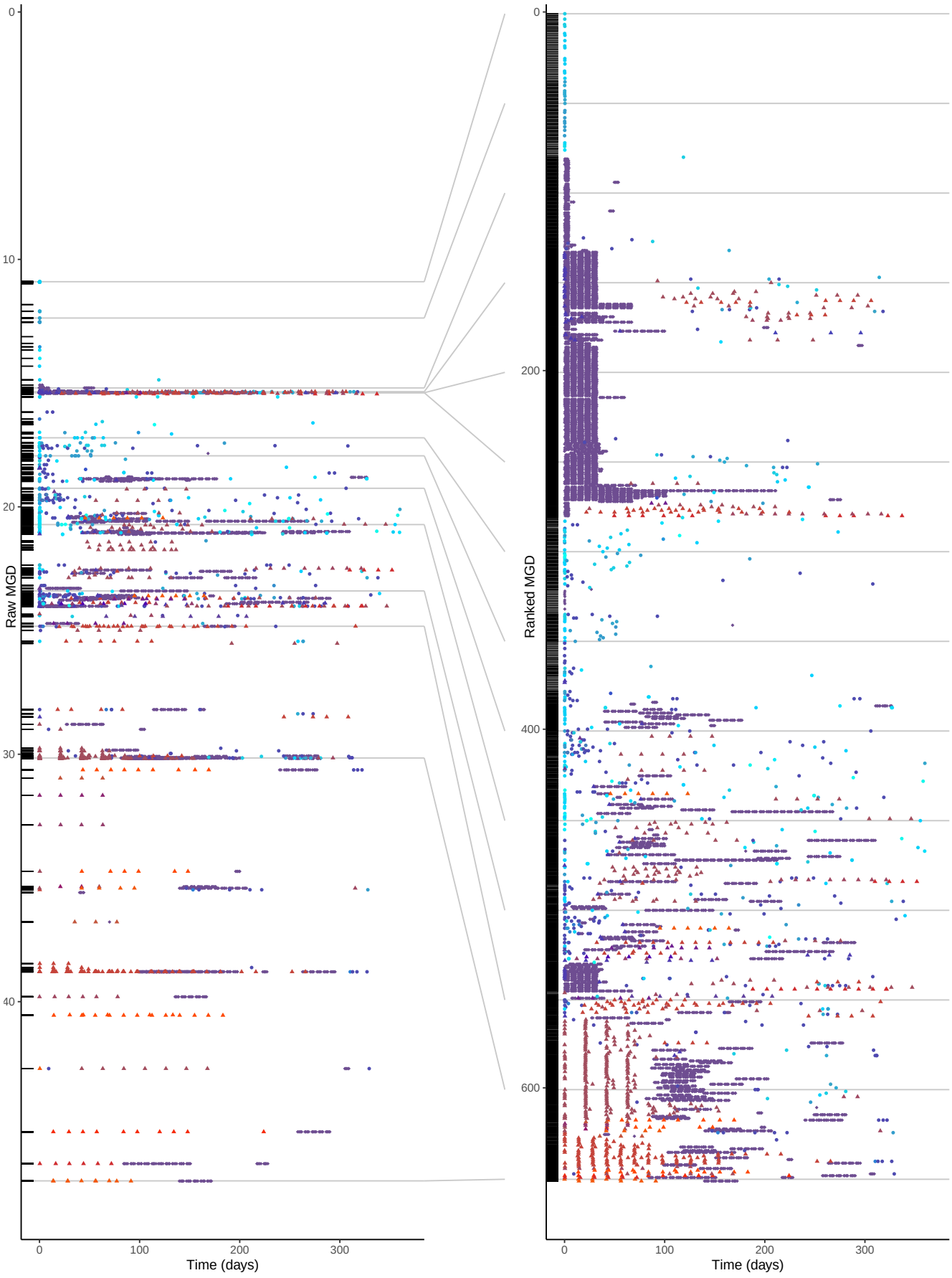

Figure 5.4. Rectal cancer pathways ordered from most to least similar to guidelines, distributed according to the raw MGD score (left) and their ranked MGD score (right)

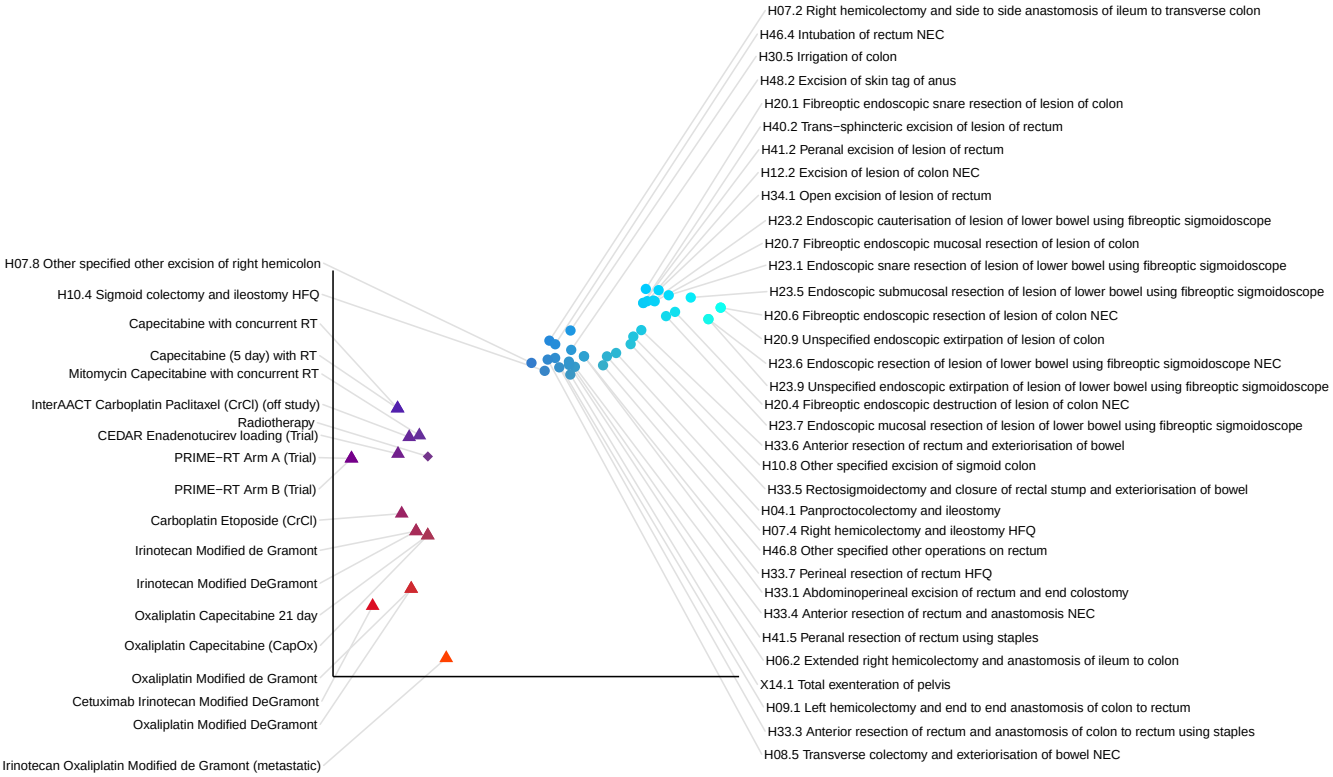

Figure 5.5. Key to OPCS concepts used in Figure 5.4

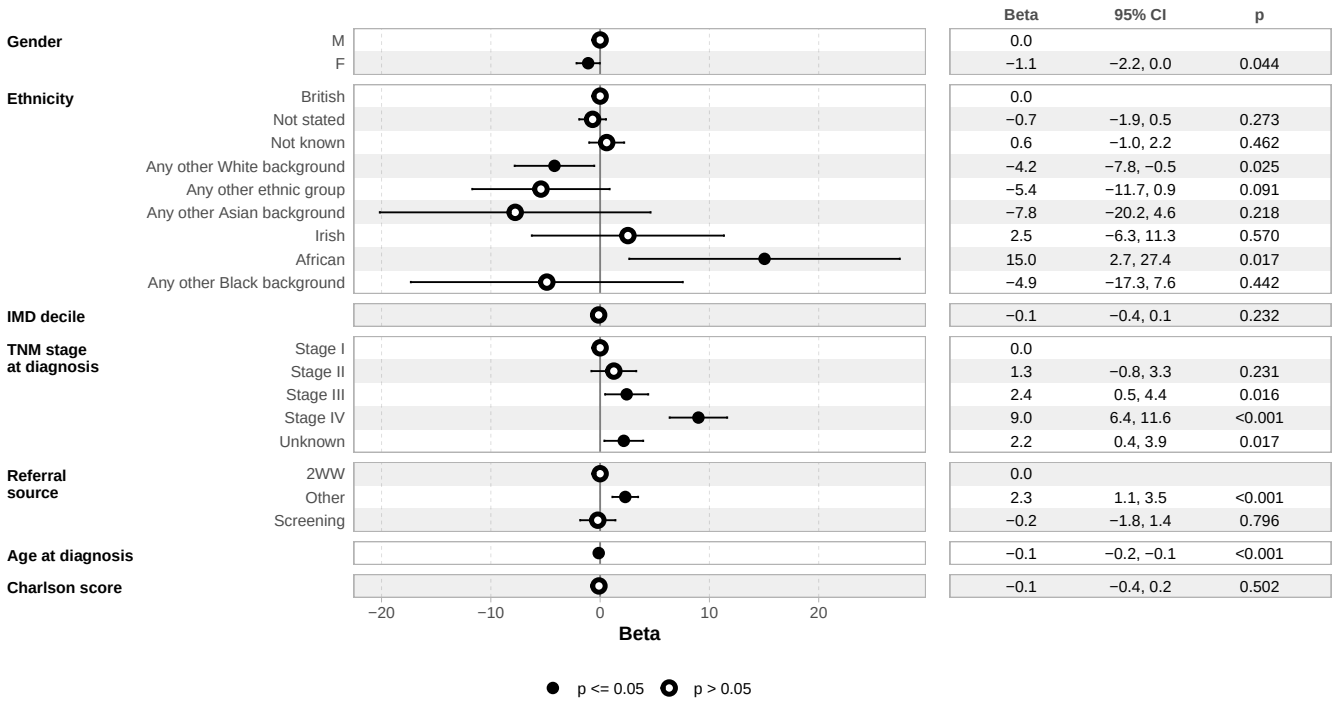

Figure 5.6. Results of a linear regression model, fitting MGD against patients' demographic attributes

the first day of every week, and sometimes involved continuous chemotherapy throughout, depending on the particular drug used. Chemoradiotherapy was therefore represented as a block of radiotherapy with chemotherapy embeddings substituted at random points, at a 3:1 ratio.

Figure 5.4 shows the traces for patients with rectal cancer, ranked by their minimum guideline distance. Relatively simple surgery- or radiotherapy-only pathways received small distance scores, followed by a large group of surgery-then-radiotherapy pathways. All of these groups cluster relatively close to each other, with E-DTW scores less than 25. Beyond this, there are a number of traces that consist of either chemotherapy alone, or chemotherapy followed by radiotherapy, which is not a possible pathway specified in the guideline model.

In order to examine patient-level factors which are associated with the MGD score, a linear regression model was fit using MGD as the dependent variable, and gender, ethnicity, Index of Multiple Deprivation (IMD), stage, referral source, age, and Charlson score as regressors (Figure 5.6). Female patients were associated with a very slightly lower MGD than male patients, whilst referrals from sources other than the main 2WW and screening routes slightly increased MGD. Age at diagnosis and Charlson score had very small negative and positive impacts on MGD respectively. Later cancer stage consistently increased MGD, although this should be interpreted with caution given the relatively high levels of patients with unknown staging. The relatively small sample sizes available for ethnicities other than "British" meant that confidence intervals for ethnicity coefficients are extremely wide, and therefore not necessarily reliable.

5.5 *Discussion*

This chapter has identified a set of key requirements for a notation that described clinical pathways, based on the actual content of UK guidelines. An examination of the existing literature on decision support languages (CIGs) and on process notations (BPMN and UML) found that they did not on the whole meet these requirements.

The three-step method presented in Section 5.3 is designed to address these. By mapping the dataset directly to the events specified in major guidelines, I ensure that the method supports the description of events at the correct level of granularity for pathway use cases. By breaking recommendations down into simple transitions, I encourage the composition of pathway models from multiple statements and guidelines. By modifying UML transitions to add maximum and minimum timing constraints, support

is introduced for the description of timing constraints between events, directly addressing the sorts of time constraints that are of interest in pathway research. Finally, I outline a method that transforms this representation into a form compatible with the E-DTW method outlined in Chapter 4. Importantly, this allows guideline-pathway distance ("compliance") to be measured as a continuous distance rather than a binary variable, which is a useful and flexible measure that supports many analyses and applications.

These results from the application of this method to rectal cancer indicate that the MGD measure applied over our representations is a useful measure that does encode similarity to guidelines. The distribution of patient traces in Figure 5.4 aligns with expected ideas around pathways, and the results of the regression analysis demonstrate its utility in identifying factors associated with guideline adherence. Disease stage was, by this measure, the most significant predictor of distance from treatment guidelines, suggesting that non-guideline-compliant treatments (chemotherapy or chemotherapy and radiotherapy without surgery) were associated with more advanced disease; age and comorbidity had relatively small effects once this was accounted for.

6 *Conclusions*

In exploring the structure, semantics, and usage of information encoded in the electronic health record, this thesis has made several novel contributions to the study of patient pathways from retrospective health data.

Pathways need context

In Chapter 3, I observed that simply applying process mining methods to data out of the box does not guarantee meaningful insights, and demonstrated that there exist demographic and clinical factors outside of the pure event data that significantly impact the interpretation of pathway data. For the field of process mining to make meaningful observations and improvements around real clinical practice, it needs to consider this deep context and integrate it into any analysis.

Ontologies can guide data preparation

In Chapter 3, I also described a method for filtering complex patient histories that makes use of relational knowledge from ontologies. The proposed method is capable of generating a reasonably complete process model, albeit with a significant trade-off between precision and recall, and provides a promising foundation for rapidly iterating over new research ideas and supporting conversations with clinicians. It has several interesting features: long codelists can be described with a small number of intuitive constraints, and it accepts and outputs concepts in standard ontologies, encouraging standardisation and sharing of codelists. Additionally, the preparation of codelists through rule-based reasoning is a useful and informative exercise: the presence and absence of certain concepts highlights the omissions and assumptions present in ontologies. Ontologies are therefore a useful framework with which to provide useful context needed for pathway analysis in a structured way, although attention does need to be paid to their original intention and design decisions to ensure that their purpose is aligned with the desired research.

KGEs encode semantic similarity

I evaluated several KGE models on the specific task of learning representations of the SNOMED CT ontology, and evaluated them using a combination of traditional link prediction metrics, and on more recent graph-based metrics. RotatE was the best-performing model according to conventional metrics, but it also demonstrated strong performance on the graph metrics, indicating that KGE representations did align with pre-existing domain knowledge and encode information in a semantically meaningful way. The success of these learned representations in the subsequent section is evidence that the learned representations are suitable for downstream uses.

Whilst a variety of representation learning algorithms are available today, knowledge graph embeddings have a number of practical advantages for health data: learning representations solely based on established ontological knowledge creates a well-founded model shareable between different providers, and a model that can be re-used for different purposes, significantly lowering the computational and environmental costs associated with modern machine learning. Creating vector representations of concepts is a useful tool that allows many different methods to be applied on top of them, and basing these representations on ontologies means that structured, established contextual knowledge can be incorporated into the learning process.

A new similarity measure combines concept embeddings with dynamic time warping

I proposed embedding-based dynamic time warping (E-DTW), a novel distance measure designed to measure the similarity between patients' treatment pathways. A thorough evaluation of eight different configurations indicated that this metric takes a number of useful and informative features into consideration, including pathway timespan and number of events, as well as the timings between events and — crucially — the semantic similarity between them. Combining KGE-generated concept embeddings with the dynamic time warping algorithm is an effective method of measuring semantic similarity between pathways, and captures both semantic differences in treatment concepts and differences in timing patterns.

Insights into the gap between guidelines and data

I then proposed an extension of the E-DTW method for measuring the distance between patient histories and recommended pathways. This method incorporates a number of steps for ensuring alignment between a particular

guideline and a dataset, a UML-based notation for representing guideline recommendations, and an algorithm that generates embeddings of hypothetical guideline-compliant patients and compares them to real histories. Through applying this method to actual rectal cancer guidelines, I identified a gap between the recommendations laid out in guidelines and the data available, indicating that in many cases only a small proportion of recommendations are measurable with current datasets. The E-DTW method is however effective at describing similarity between pathways and guidelines.

What all of these contributions have in common is that they explore the fundamental semantic meaning of structured health data, in terms of its implicit ontological knowledge and the process knowledge that produces it, and exploit this in order to develop a deeper understanding of pathways. Pathways and data are strongly influence by a great deal of surrounding contextual factors, and I have demonstrated that ontological knowledge can provide a framework for integrating this context into research. Furthermore, this ontological knowledge can be effectively represented in numerical forms, and used as input for machine learning algorithms, meaning that this established structural information can be injected into methods that are otherwise purely statistical, allowing the benefits of both approaches to be used. Issues around the interpretation and study of patient pathways arise from treating them as strictly deterministic and structured; a promising solution to this is more flexible analysis methods that create a view of the patient pathway grounded in *similarity* and *distance* rather than rigid *compliance*, which I have provided in the form of E-DTW.

There are several key learnings that can be taken from this work. For practitioners of process mining, care needs to be taken to consider context and sub-populations in analysis, and to document and justify event concepts that are included and included for analysis. From the point of view of representation learning and AI, a model does not need to score exceptionally highly on the traditional metrics to encode useful structural information by other metrics, or to be a useful input for downstream applications. However, all users of structured domain knowledge need to consider the original intent and purpose of *both* data and ontologies, to ensure that they are aligned with research objectives; when it comes to integrating structure and context into analysis, quality is better than quantity.

6.1 *Limitations and future work*

The method for ontological-informed filtering of event logs described in Chapter 4 varies significantly in its effectiveness in different disease areas. A major factor in this was the different codelists used as benchmarks: colorectal cancer showed the best performance, but it also had by far the most detailed and reliable codelist, specifically designed for investigating different treatments, whilst in other disease areas the best list available was often designed for different purposes. A more effective evaluation might therefore make use of equally thorough codelists across all disease areas; this speaks to more widespread challenge across medical research, but with time and funding the development of more rigorous codelists would be possible. With enough time and resources, it would be possible to iteratively develop an equally thorough codelist for the other diseases by conversation and consensus amongst domain experts; repeating this process for four diseases was not feasible within this project.

The codelists generated by this approach were far from perfect: their shortcomings emphasise the importance of conducting research in partnership with clinicians and developing established standard lists of concepts. It is likely that the ideal solution exists by hybridising ontological, statistical, and human approaches: this has the potential to be an extremely fruitful area of future research. Modern, large-scale machine learning methods such as LLMs have largely been avoided in this thesis, as part of a deliberate decision to deal with established and quality-assured expert data, but such models could make useful contributions in examining the literature on an extremely large scale to identify potential associations to explore in depth.

Chapter 4 focused on using knowledge graph embeddings as a method for learning representations, which were specifically chosen to address a number of practical issues and ensure that the proposed approach remained reproducible and accessible. At the time of writing, research in representation learning, in particular deep learning, is proceeding at pace; these methods could well produce more effective and semantically meaningful representations of medical concepts. However, this is not a guarantee. Previous studies have indicated that large-scale, general-purpose models are often outperformed by domain-specific models in medicine. Chapter 4 proves that KGE-generated representations are good enough for the purposes of the E-DTW algorithm; any proposal to replace them with more advanced or complex models needs to justify itself against these practical concerns.

Chapter 5 highlighted the existence of a gap between the recommendations laid down in guidelines, and the ability of many research datasets to

actually monitor them. Some of this data is just not recorded at all; some is recorded, but not routinely published or extracted as part of research datasets. In the case of the latter, the increased investment in the UK's health data research infrastructure and the increasing availability of new health data resources should be considered an opportunity to create datasets that are more closely aligned with the needs of research questions. In the longer term, a significant area for future research should be the publication of clinical guidelines in computable form. This is a longstanding research goal — in particular in the CIG literature — and more recently, NICE have made clear their intentions to advance it (Scott et al. 2023), but the focus remains on decision support. More holistic analyses of patients' whole pathways have a distinct set of requirements: ideally, developers of computable guidelines should also take these into account.

The E-DTW method and its extensions described in this thesis has delivered promising results in the area of colorectal cancer, a major disease with real research needs surrounding pathways: the next step is to evaluate its effectiveness in different disease areas. The methods should also be tested on larger regional and national datasets, in order to get the most meaningful insights. Whilst the results in this thesis exist for the demonstration of the algorithms, and caution is advised when attempting to read epidemiological meaning into their results, it is clear that sample size and data quality are problems when trying to investigate pathway differences. Gathering more widespread data and testing these methods on different populations should be considered a high priority if we want these methods to realise their potential in monitoring, identifying, and addressing inequalities in healthcare provision and outcomes.

In Chapter 4, it was relatively hard to observe patterns relating to deprivation in the clusters due to the relative homogeneity of the sample, being from one geographic area. Similarly, it was not possible to obtain a useful estimate of the effect of ethnicity on pathways in Chapter 5 due to very small sample sizes available for many groups. A wider deployment of this approach would help to generate insights into pathway variations in these populations. Research indicates that there are real differences in the quality of care offered to people depending on their background; effective pathway analysis tools, applied over large-scale informative datasets, have the potential to deliver insights into these patterns, and ultimately help to rectify these inequities.

6.2 *Closing remarks*

The increasing availability of health and care data for research purposes opens up many exciting possibilities for new insights into disease and treatment, but this data needs effective computational methods to derive these insights. This thesis has focused on the data itself: its meaning, its structure, and its limitations, in order to develop a realistic idea of what is possible. Structured data is, by virtue of the way it is recorded and encoded, already imbued with incredibly rich semantics that can be exploited to create useful insights.

Pathways are inherent complex, and purely logical or statistical attempts to analyse them have drawbacks. This thesis has attempted to chart a new path: using modern and flexible machine learning concepts such as similarity measures and clustering that allow pathways to be analysed as points on a continuous spectrum, whilst also grounding this learning and taking advantage of well-founded, expert knowledge in ontologies. It is this interplay between sub-symbolic representation and logic that will form the basis of the most exciting advances in data in the future, allowing the benefits of health data to be realised whilst also ensuring analysis is relevant, reliable, and reproducible.

Bibliography

Agar, Jon (Sept. 2020). "What is science for? The Lighthill report on artificial intelligence reinterpreted". In: *The British Journal for the History of Science* 53.3, pp. 289–310. DOI: 10.1017/S0007087420000230.

Aggarwal, Ajay et al. (July 1, 2022). "What really matters for cancer care — health systems strengthening or technological innovation?" In: *Clinical Oncology* 34.7, pp. 430–435. DOI: 10.1016/j.clon.2022.02.012.

Aggarwal, Ajay et al. (July 8, 2024). "NHS cancer services and systems — ten pressure points a UK cancer control plan needs to address". In: *The Lancet Oncology*. DOI: 10.1016/S1470-2045(24)00345-0.

Ali, Mehdi et al. (2021). "PyKEEN 1.0: a Python library for training and evaluating knowledge graph embeddings". In: *Journal of Machine Learning Research* 22.82, pp. 1–6. ISSN: 1533-7928.

Ali, Mehdi et al. (Dec. 2022). "Bringing light into the dark: a large-scale evaluation of knowledge graph embedding models under a unified framework". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 44.12, pp. 8825–8845. DOI: 10.1109/TPAMI.2021.3124805.

Allemani, Claudia et al. (Mar. 17, 2018). "Global surveillance of trends in cancer survival 2000–14 (CONCORD-3): analysis of individual records for 37 513 025 patients diagnosed with one of 18 cancers from 322 population-based registries in 71 countries". In: *The Lancet* 391.10125, pp. 1023–1075. DOI: 10.1016/S0140-6736(17)33326-3.

Allen, Davina, Elizabeth Gillen & Laura Rixson (June 2009). "Systematic review of the effectiveness of integrated care pathways: what works, for whom, in which circumstances?" In: *International Journal of Evidence-Based Healthcare* 7.2, pp. 61–74. DOI: 10.1111/j.1744-1609.2009.00127.x.

Alves de Medeiros, Ana Karla & Wil M. P. van der Aalst (2009). "Process mining towards semantics". In: *Advances in Web Semantics I: Ontologies, Web Services and Applied Semantic Web*. Ed. by Tharam S. Dillon et al. Lecture Notes in Computer Science. Berlin & Heidelberg, DE: Springer, pp. 35–80. DOI: 10.1007/978-3-540-89784-2_3.

Andaur Navarro, Constanza L. et al. (June 1, 2023). "Systematic review finds "spin" practices and poor reporting standards in studies on machine

learning-based prediction models". In: *Journal of Clinical Epidemiology* 158, pp. 99–110. DOI: 10.1016/j.jclinepi.2023.03.024.

Argilés, G. et al. (Oct. 1, 2020). "Localised colon cancer: ESMO Clinical Practice Guidelines for diagnosis, treatment and follow-up". In: *Annals of Oncology* 31.10, pp. 1291–1305. DOI: 10.1016/j.annonc.2020.06.022.

Aspland, Emma et al. (Mar. 2021). "Modified Needleman-Wunsch algorithm for clinical pathway clustering". In: *Journal of Biomedical Informatics* 115, p. 103668. DOI: 10.1016/j.jbi.2020.103668.

Atolagbe, Oluseun O. et al. (Sept. 27, 2024). "Coding rules for uncertain and "ruled out" diagnoses in ICD-10 and ICD-11". In: *BMC Medical Informatics and Decision Making* 21.6, p. 386. DOI: 10.1186/s12911-024-02661-6.

Bagnall, Anthony et al. (May 1, 2017). "The great time series classification bake off: a review and experimental evaluation of recent algorithmic advances". In: *Data Mining and Knowledge Discovery* 31.3, pp. 606–660. DOI: 10.1007/s10618-016-0483-9.

Bainbridge, Michael (2019). "Big data challenges for clinical and precision medicine". In: *Big Data, Big Challenges: A Healthcare Perspective: Background, Issues, Solutions and Research Directions*. Ed. by Mowafa Househ, Andre W. Kushniruk & Elizabeth M. Borycki. Lecture Notes in Bioengineering. Cham, CH: Springer International, pp. 17–31. DOI: 10.1007/978-3-030-06109-8_2.

Banda, Juan M. et al. (July 2018). "Advances in electronic phenotyping: from rule-based definitions to machine learning models". In: *Annual Review of Biomedical Data Science* 1, p. 53. DOI: 10.1146/annurev-biodatasci-080917-013315.

Barban, Nicola & Francesco C. Billari (2012). "Classifying life course trajectories: a comparison of latent class and sequence analysis". In: *Journal of the Royal Statistical Society, Series C (Applied Statistics)* 61.5, pp. 765–784. DOI: 10.1111/j.1467-9876.2012.01047.x.

BASO (Apr. 9, 2020). *BASO Guidance - Strategy for cancer surgery sustainability and recovery in the COVID 19 pandemic*. British Association of Surgical Oncology ~ The Association for Cancer Surgery. URL: https://baso.org.uk/media/99217/baso_guidance_for_cancer_surgery_9th_april_2020_v7.pdf.

Beam, Andrew L. et al. (2020). "Clinical concept embeddings learned from massive sources of multimodal medical data". In: *Biocomputing 2020*. Pacific Symposium on Biocomputing. Vol. 25. Kohala Coast, HI, USA, pp. 295–306. DOI: 10.1142/9789811215636_0027.

Bean, Daniel M. et al. (Nov. 27, 2017). "Knowledge graph prediction of unknown adverse drug reactions and validation in electronic health records". In: *Scientific Reports* 7.1, p. 16416. DOI: 10.1038/s41598-017-16674-x.

Bellomarini, Luigi, Emanuel Sallinger & Sahar Vahdati (2020). "Reasoning in knowledge graphs: an embeddings spotlight". In: *Knowledge Graphs and Big Data Processing*. Ed. by Valentina Janev et al. Lecture Notes in Computer Science. Springer, pp. 87–101. DOI: 10.1007/978-3-030-53199-7_6.

Berndt, Donald J. & James Clifford (July 31, 1994). "Using dynamic time warping to find patterns in time series". In: *Proceedings of the 3rd International Conference on Knowledge Discovery and Data Mining*. KDD 1994. Seattle, WA, USA, pp. 359–370. DOI: 10.5555/3000850.3000887.

Berti, Alessandro et al. (July 20, 2023). "Graph-based feature extraction on object-centric event logs". In: *International Journal of Data Science and Analytics*. DOI: 10.1007/s41060-023-00428-2.

Bertolini, Lorenzo S. (May 26, 2023). "Assessing compositionality with phrase-level adjective-noun entailment". Doctoral thesis. University of Sussex. URL: https://hdl.handle.net/10779/uos.23496527.v1 (retrieved 04/09/2024).

Bettencourt-Silva, Joao H. et al. (July 10, 2015). "Building data-driven pathways from routinely collected hospital data: a case study on prostate cancer". In: *JMIR Medical Informatics* 3.3, e26. DOI: 10.2196/medinform.4221.

Bharadhwaj, Vinay Srinivas et al. (Oct. 1, 2021). "CLEP: a hybrid data- and knowledge-driven framework for generating patient representations". In: *Bioinformatics* 37.19, pp. 3311–3318. DOI: 10.1093/bioinformatics/btab340.

Bhavani, Sivasubramanium V et al. (June 1, 2023). "Comparison of time series clustering methods for identifying novel subphenotypes of patients with infection". In: *Journal of the American Medical Informatics Association* 30.6, pp. 1158–1166. DOI: 10.1093/jamia/ocad063.

Biggin, Fran et al. (Nov. 6, 2023). "Discovering patterns in outpatient neurology appointments using state sequence analysis". In: *BMC Health Services Research* 23.1, p. 1208. DOI: 10.1186/s12913-023-10218-y.

Bion, Julian et al. (July 9, 2021). "Increasing specialist intensity at weekends to improve outcomes for patients undergoing emergency hospital admission: the HiSLAC two-phase mixed-methods study". In: *Health Services and Delivery Research* 9.13, pp. 1–166. DOI: 10.3310/hsdr09130.

Biswas, Russa et al. (2023). "Knowledge graph embeddings: open challenges and opportunities". In: *Transactions on Graph Data and Knowledge* 1.1, 4:1–4:32. DOI: 10.4230/TGDK.1.1.4.

Bordes, Antoine et al. (Dec. 2013). "Translating embeddings for modeling multi-relational data". In: *Advances in Neural Information Processing Systems*. NeurIPS 2013. Ed. by Burges, CJ et al. Vol. 26. Lake Tahoe, NV, USA: Curran Associates, Inc., pp. 2787–2795. ISBN: 978-1-63266-024-4.

Börger, Egon (July 1, 2012). "Approaches to modeling business processes: a critical analysis of BPMN, workflow patterns and YAWL". In: *Software & Systems Modeling* 11.3, pp. 305–318. DOI: 10.1007/s10270-011-0214-z.

Boxwala, Aziz A et al. (Dec. 1, 2011). "A multi-layered framework for disseminating knowledge for computer-based decision support". In: *Journal of the American Medical Informatics Association* 18 (Supplement 1), pp. i132–i139. DOI: 10.1136/amiajnl-2011-000334.

Boyle, Jemma M. et al. (May 1, 2020). "Determinants of variation in the use of adjuvant chemotherapy for stage III colon cancer in England". In: *Clinical Oncology* 32.5, e135–e144. DOI: 10.1016/j.clon.2019.12.008.

Braun, Richard et al. (Nov. 2014). "BPMN4CP: Design and implementation of a BPMN extension for clinical pathways". In: *IEEE International Conference on Bioinformatics and Biomedicine*. BIBM 2014. Belfast, UK: IEEE, pp. 9–16. DOI: 10.1109/BIBM.2014.6999261.

Brown, Katherine E et al. (May 1, 2025). "Large language models are less effective at clinical prediction tasks than locally trained machine learning models". In: *Journal of the American Medical Informatics Association* 32.5, pp. 811–822. DOI: 10.1093/jamia/ocaf038.

Burke, Hannah et al. (Feb. 1, 2022). "Biomarker identification using dynamic time warping analysis: a longitudinal cohort study of patients with COVID-19 in a UK tertiary hospital". In: *BMJ Open* 12.2, e050331. DOI: 10.1136/bmjopen-2021-050331.

Campbell, W. Scott et al. (Oct. 1, 2015). "An alternative database approach for management of SNOMED CT and improved patient data queries". In: *Journal of Biomedical Informatics* 57, pp. 350–357. DOI: 10.1016/j.jbi.2015.08.016.

Cancer Research UK (May 14, 2015). *Bowel cancer statistics*. Cancer Research UK. URL: https://www.cancerresearchuk.org/health-professional/cancer-statistics/statistics-by-cancer-type/bowel-cancer (retrieved 05/12/2022).

Carr, Oliver et al. (Nov. 28, 2021). "Longitudinal patient stratification of electronic health records with flexible adjustment for clinical outcomes". In: *Proceedings of Machine Learning for Health*. ML4H 2021. Vol. 158.

Proceedings of Machine Learning Research, pp. 220–238. URL: https://proceedings.mlr.press/v158/carr21a.html (retrieved 05/10/2022).

Carvalho, Ricardo M.S., Daniela Oliveira & Catia Pesquita (Jan. 19, 2023). "Knowledge graph embeddings for ICU readmission prediction". In: *BMC Medical Informatics and Decision Making* 23, p. 12. DOI: 10.1186/s12911-022-02070-7.

Castela Forte, José et al. (June 8, 2021). "Identifying and characterizing high-risk clusters in a heterogeneous ICU population with deep embedded clustering". In: *Scientific Reports* 11.1, p. 12109. DOI: 10.1038/s41598-021-91297-x.

Chae, Junghoon et al. (July 2020). "Converting clinical pathways to BPM+ standards: a case study in stable ischemic heart disease". In: *IEEE 33rd International Symposium on Computer-Based Medical Systems*. CBMS 2020. Rochester, MN, USA, pp. 453–456. DOI: 10.1109/CBMS49503.2020.00092.

Chammas, Lara et al. (Jan. 3, 2024). "Care records and healthcare processes: adding context to clinical codes". In: *Proceedings of the 57th Hawaii International Conference on System Sciences*. HICSS 2024. Honolulu, HI, USA: University of Hawai'i at Mānoa, pp. 3697–3706. ISBN: 978-0-9981331-7-1.

Chang, David et al. (July 2020). "Benchmark and best practices for biomedical knowledge graph embeddings". In: *Proceedings of the 19th SIGBioMed Workshop on Biomedical Language Processing*. BioNLP 2020. Online, pp. 167–176. DOI: 10.18653/v1/2020.bionlp-1.18.

Charlson, Mary E. et al. (Jan. 1, 1987). "A new method of classifying prognostic comorbidity in longitudinal studies: development and validation". In: *Journal of Chronic Diseases* 40.5, pp. 373–383. DOI: 10.1016/0021-9681(87)90171-8.

Cheikhrouhou, Saoussen et al. (Dec. 2, 2013). "Toward a time-centric modeling of business processes in BPMN 2.0". In: *Proceedings of the 15th International Conference on Information Integration and Web-based Applications & Services*. IIWAS 2013. Vienna, AT, pp. 154–163. DOI: 10.1145/2539150.2539182.

Chen, Jiaoyan et al. (2023). "Knowledge graphs for the life sciences: recent developments, challenges and opportunities". In: *Transactions on Graph Data and Knowledge* 1.1, 5:1–5:33. DOI: 10.4230/TGDK.1.1.5.

Chen, Shan et al. (Apr. 1, 2024). "Evaluating the ChatGPT family of models for biomedical reasoning and classification". In: *Journal of the American Medical Informatics Association* 31.4, pp. 940–948. DOI: 10.1093/jamia/ocad256.

Choi, Edward et al. (Aug. 13, 2016). "Multi-layer representation learning for medical concepts". In: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. KDD 2016. New York, NY, USA: Association for Computing Machinery, pp. 1495–1504. DOI: 10.1145/2939672.2939823.

Coleman, M. P. et al. (Jan. 8, 2011). "Cancer survival in Australia, Canada, Denmark, Norway, Sweden, and the UK, 1995–2007 (the International Cancer Benchmarking Partnership): an analysis of population-based cancer registry data". In: *The Lancet* 377.9760, pp. 127–138. DOI: 10.1016/S0140-6736(10)62231-3.

Cordeschi, Roberto (Apr. 25, 2007). "AI turns fifty: revisiting its origins". In: *Applied Artificial Intelligence* 21.4–5, pp. 259–279. DOI: 10.1080/08839510701252304.

CORECT-R Data Coding (Oct. 2020). UK Colorectal Cancer Intelligence Hub. URL: https://www.ndph.ox.ac.uk/corectr/files/corect-r-data-coding-v1-0-oct20.pdf.

Cremerius, Jonas et al. (2023). "Event log generation in MIMIC-IV". In: *Process Mining Workshops*. Ed. by Marco Montali, Arik Senderovich & Matthias Weidlich. Lecture Notes in Business Information Processing. Cham, CH: Springer Nature, pp. 302–314. DOI: 10.1007/978-3-031-27815-0_22.

d'Amato, Claudia et al. (Dec. 19, 2023). "Machine learning and knowledge graphs: existing gaps and future research challenges". In: *Transactions on Graph Data and Knowledge* 1.1, 8:1–8:35. DOI: 10.4230/TGDK.1.1.8.

Dagliati, A. et al. (Feb. 1, 2017). "Temporal electronic phenotyping by mining careflows of breast cancer patients". In: *Journal of Biomedical Informatics* 66, pp. 136–147. DOI: 10.1016/j.jbi.2016.12.012.

Damerau, Fred J. (Mar. 1, 1964). "A technique for computer detection and correction of spelling errors". In: *Communications of the ACM* 7.3, pp. 171–176. DOI: 10.1145/363958.363994.

De Bleser, Leentje et al. (2006). "Defining pathways". In: *Journal of Nursing Management* 14.7, pp. 553–563. DOI: 10.1111/j.1365-2934.2006.00702.x.

De Roock, Emmelien & Niels Martin (Mar. 1, 2022). "Process mining in healthcare – an updated perspective on the state of the art". In: *Journal of Biomedical Informatics* 127, p. 103995. DOI: 10.1016/j.jbi.2022.103995.

Denaxas, Spiros et al. (Dec. 1, 2019). "UK phenomics platform for developing and validating electronic health record phenotypes: CALIBER". In: *Journal of the American Medical Informatics Association* 26.12, pp. 1545–1559. DOI: 10.1093/jamia/ocz105.

Department of Health and Social Care (Dec. 23, 2022). *Secure data environment for NHS health and social care data - policy guidelines*. URL: https://www.gov.uk/government/publications/secure-data-environment-policy-guidelines/secure-data-environment-for-nhs-health-and-social-care-data-policy-guidelines (retrieved 10/18/2024).

— (June 23, 2023). *£21 million to roll out artificial intelligence across the NHS*. URL: https://www.gov.uk/government/news/21-million-to-roll-out-artificial-intelligence-across-the-nhs (retrieved 07/16/2024).

Dhiman, Paula et al. (May 1, 2023). "Overinterpretation of findings in machine learning prediction model studies in oncology: a systematic review". In: *Journal of Clinical Epidemiology* 157, pp. 120–133. DOI: 10.1016/j.jclinepi.2023.03.012.

Doo, Florence X. et al. (Feb. 2024). "Environmental sustainability and AI in radiology: a double-edged sword". In: *Radiology* 310.2, e232030. DOI: 10.1148/radiol.232030.

Downing, Amy et al. (Oct. 1, 2021). "Data resource profile: The COloREC-Tal cancer data repository (CORECT-R)". In: *International Journal of Epidemiology* 50.5, 1418–1418k. DOI: 10.1093/ije/dyab122.

Drosdowsky, Allison & Karla Gough (Aug. 1, 2022). "The Charlson Comorbidity Index: problems with use in epidemiological research". In: *Journal of Clinical Epidemiology* 148, pp. 174–177. DOI: 10.1016/j.jclinepi.2022.03.022.

Dwyer, Owen P. et al. (Sept. 2023). "Reasoning over health records with Vadalog: a rule-based approach to patient pathways". In: *Proceedings of the 17th International Rule Challenge and 7th Doctoral Consortium*. RuleML+RR 2023. Ed. by Jan Vanthienen et al. Vol. 3485. CEUR Workshop Proceedings. Oslo, NO. URL: https://ceur-ws.org/Vol-3485/.

Dwyer, Owen P. et al. (2024). "Investigating an ontology-informed approach to event log generation in healthcare". In: *Process Mining Workshops*. ICPM 2023. Ed. by Johannes De Smedt & Pnina Soffer. Vol. 503. Lecture Notes in Business Information Processing. Cham, CH: Springer, pp. 71–83. DOI: 10.1007/978-3-031-56107-8_18.

— (May 29, 2025). "Using ontologies to facilitate healthcare process mining and analysis". In: *Journal of Intelligent Information Systems*. DOI: 10.1007/s10844-025-00942-8.

Ehrlinger, Lisa & Wolfram Wöß (2016). "Towards a definition of knowledge graphs". In: *Joint Proceedings of the Posters and Demos Track of the 12th International Conference on Semantic Systems*. SEMANTiCS 2016. Ed. by

Michael Martin, Martí Cuquet & Erwin Folmer. Vol. 1695. Leipzig, DE. URL: https://ceur-ws.org/Vol-1695/paper4.pdf.

Elkheder, Musaab et al. (Feb. 1, 2023). "Translating and evaluating historic phenotyping algorithms using SNOMED CT". In: *Journal of the American Medical Informatics Association* 30.2, pp. 222–232. DOI: 10.1093/jamia/ocac158.

Emamjome, Fahame et al. (June 15, 2020). "Alohomora: Unlocking data quality causes through event log context". In: *Proceedings of the 28th European Conference on Information Systems*. ECIS 2020. Online. ISBN: 978-1-7336325-1-5.

Fawcett, Nicola et al. (Sept. 4, 2019). "'Caveat emptor': the cautionary tale of endocarditis and the potential pitfalls of clinical coding data-an electronic health records study". In: *BMC Medicine* 17.1, p. 169. DOI: 10.1186/s12916-019-1390-x.

Fazakarley, C.A. et al. (Dec. 1, 2023). "Experiences of using artificial intelligence in healthcare: a qualitative study of UK clinician and key stakeholder perspectives". In: *BMJ Open* 13.12, e076950. DOI: 10.1136/bmjopen-2023-076950.

Fearnhead, Nicola et al. (Apr. 9, 2020). *Joint ACPGBI, BSG and BSGAR considerations for adapting the rapid access colorectal cancer pathway during COVID-19 pandemic*. Association of Coloproctology of Great Britain and Ireland. URL: https://www.acpgbi.org.uk/about/news/381/joint_acpgbi_bsg_and_bsgar_considerations_for_adapting_the_rapid_access_colorectal_cancer_pathway_during_covid19_pandemic/ (retrieved 08/13/2024).

Fenton, Hayley M. et al. (2021). "National variation in pulmonary metastasectomy for colorectal cancer". In: *Colorectal Disease* 23.6, pp. 1306–1316. DOI: 10.1111/codi.15506.

Floridi, Luciano (Nov. 5, 2024). "Why the AI hype is another tech bubble". In: *Philosophy & Technology* 37.4, p. 128. DOI: 10.1007/s13347-024-00817-w.

Forster, K. et al. (Feb. 2020). "Can concordance between actual care received and a pathway map be measured on a population level in Ontario? A pilot study". In: *Current Oncology* 27.1, e27–e33. DOI: 10.3747/co.27.5349.

Freemantle, Nick et al. (Sept. 5, 2015). "Increased mortality associated with weekend hospital admission: a case for expanded seven day services?" In: *BMJ* 351, h4596. DOI: 10.1136/bmj.h4596.

Fu, Mingzhou et al. (Mar. 1, 2023). "Defining the distance between diseases using SNOMED CT embeddings". In: *Journal of Biomedical Informatics* 139, p. 104307. DOI: 10.1016/j.jbi.2023.104307.

Gagne, Denis & André Trudel (July 2009). "Time-BPMN". In: *IEEE Conference on Commerce and Enterprise Computing*. CEC 2009. Vienna, AT, pp. 361–367. DOI: 10.1109/CEC.2009.71.

Gamba, Magdalena (2017). "Guideline representation and execution tools: an evaluation study". Master's thesis. Universiteit van Amsterdam. URL: https://scripties.uba.uva.nl/search?id=636152.

Garbi, Madalina (June 1, 2021). "National Institute for Health and Care Excellence clinical guidelines development principles and processes". In: *Heart* 107.12, pp. 949–953. DOI: 10.1136/heartjnl-2020-318661.

Gatta, Roberto et al. (Dec. 4, 2017a). "Generating and comparing knowledge graphs of medical processes using pMineR". In: *Proceedings of the 9th Knowledge Capture Conference*. K-CAP 2017. New York, NY, USA: Association for Computing Machinery, pp. 1–4. DOI: 10.1145/3148011.3154464.

Gatta, Roberto et al. (2017b). "pMineR: an innovative R library for performing process mining in medicine". In: *Artificial Intelligence in Medicine*. AIME 2017. Ed. by Annette ten Teije et al. Vol. 10259. Lecture Notes in Computer Science. Cham, CH: Springer International, pp. 351–355. DOI: 10.1007/978-3-319-59758-4_42.

Geh, Ian et al. (June 20, 2017). "Association of Coloproctology of Great Britain & Ireland (ACPGBI): Guidelines for the Management of Cancer of the Colon, Rectum and Anus (special issue)". In: *Colorectal Disease* 19 (Supplement 1). DOI: 10.1111/codi.13697.

Gema, Aryo Pradipta et al. (Jan. 5, 2024). "Knowledge graph embeddings in the biomedical domain: are they useful? A look at link prediction, rule learning, and downstream polypharmacy tasks". In: *Bioinformatics Advances* 4.1, vbae097. DOI: 10.1093/bioadv/vbae097.

Gencturk, Mert et al. (May 27, 2024). "Transforming evidence-based clinical guidelines into implementable clinical decision support services: the CAREPATH study for multimorbidity management". In: *Frontiers in Medicine* 11. DOI: 10.3389/fmed.2024.1386689.

Genon, Nicolas, Patrick Heymans & Daniel Amyot (Oct. 12, 2010). "Analysing the cognitive effectiveness of the BPMN 2.0 visual notation". In: *Software Language Engineering: Third International Conference*. SLE 2010. Ed. by Brian Malloy, Steffen Staab & Mark van den Brand. Vol. 6563. Lecture Notes in Computer Science. Berlin & Heidelberg, DE: Springer, pp. 377–396. DOI: 10.1007/978-3-642-19440-5_25.

Gerber, Aurona & Sunet Eybers (Aug. 15, 2025). "Can LLMs reason with logic?" In: *AMCIS 2025 Proceedings*. Montréal, CA: Association for Information

Systems. URL: https://aisel.aisnet.org/amcis2025/intelfuture/intelfuture/34.

Getzen, Emily et al. (2024). "Mining for health: a comparison of word embedding methods for analysis of EHRs data". In: *Statistics in Precision Health: Theory, Methods and Applications*. Ed. by Yichuan Zhao & Ding-Geng Chen. ICSA Book Series in Statistics. Cham, CH: Springer, pp. 313–338. DOI: 10.1007/978-3-031-50690-1_13.

Giannoula, Alexia et al. (Mar. 9, 2018). "Identifying temporal patterns in patient disease trajectories using dynamic time warping: A population-based study". In: *Scientific Reports* 8.1, p. 4216. DOI: 10.1038/s41598-018-22578-1.

Giannoula, Alexia et al. (June 16, 2021). "A system-level analysis of patient disease trajectories based on clinical, phenotypic and molecular similarities". In: *Bioinformatics* 37.10, pp. 1435–1443. DOI: 10.1093/bioinformatics/btaa964.

Giannoula, Alexia et al. (Apr. 1, 2024). "Exploring long-term breast cancer survivors' care trajectories using dynamic time warping-based unsupervised clustering". In: *Journal of the American Medical Informatics Association* 31.4, pp. 820–831. DOI: 10.1093/jamia/ocad251.

Gill, T. Grandon (Mar. 1, 1995). "Early expert systems: where are they now?" In: *MIS Quarterly* 19.1, pp. 51–81. DOI: 10.2307/249711.

Glynne-Jones, R. et al. (July 1, 2017). "Rectal cancer: ESMO Clinical Practice Guidelines for diagnosis, treatment and follow-up". In: *Annals of Oncology* 28, pp. iv22–iv40. DOI: 10.1093/annonc/mdx224.

Goel, Ashok K. (2021). "Looking back, looking ahead: symbolic versus connectionist AI". In: *AI Magazine* 42.4, pp. 83–85. DOI: 10.1609/aaai.12026.

Gogleva, Anna et al. (Dec. 2022). "Knowledge graph-based recommendation framework identifies drivers of resistance in EGFR mutant non-small cell lung cancer". In: *Nature Communications* 13.1, p. 1667. DOI: 10.1038/s41467-022-29292-7.

Goldacre, Ben & Jessica Morley (Apr. 7, 2022). *Better, Broader, Safer: Using Health Data for Research and Analysis*. Department of Health and Social Care. URL: https://www.gov.uk/government/publications/better-broader-safer-using-health-data-for-research-and-analysis (retrieved 12/15/2023).

Gordon, Colin (July 1, 1996). "May we support your decision?" In: *Journal of Health Services Research & Policy* 1.3, pp. 175–178. DOI: 10.1177/135581969600100312.

Greenes, Robert A. et al. (Feb. 1, 2018). "Clinical decision support models and frameworks: seeking to address research issues underlying implemen-

tation successes and failures". In: *Journal of Biomedical Informatics* 78, pp. 134–143. DOI: 10.1016/j.jbi.2017.12.005.

Grover, Aditya & Jure Leskovec (Aug. 2016). "Node2vec: Scalable feature learning for networks". In: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. KDD 2016. New York, NY, USA, pp. 855–864. DOI: 10.1145/2939672.2939754.

Gupta, Vivek et al. (Apr. 3, 2020). "P-SIF: document embeddings using partition averaging". In: *Proceedings of the AAAI Conference on Artificial Intelligence*. AAAI 2020. Vol. 34. 05. New York, NY, USA, pp. 7863–7870. DOI: 10.1609/aaai.v34i05.6292.

Gutierrez, Claudio & Juan F. Sequeda (Feb. 22, 2021). "Knowledge graphs". In: *Communications of the ACM* 64.3, pp. 96–104. DOI: 10.1145/3418294.

Gyrard, Amelie et al. (2018). "Personalized health knowledge graph". In: *Joint Proceedings of the International Workshops on Contextualized Knowledge Graphs, and Semantic Statistics*. ISWC 2018. Ed. by Sarven Capadisli et al. Vol. 2317. CEUR Workshop Proceedings. URL: http://ceur-ws.org/Vol-2317/article-05.pdf.

Haendel, Melissa A., Christopher G. Chute & Peter N. Robinson (Oct. 11, 2018). "Classification, ontology, and precision medicine". In: *The New England Journal of Medicine* 379.15, pp. 1452–1462. DOI: 10.1056/NEJMra1615014.

Haigh, Thomas (Nov. 17, 2023). "There was no 'first AI winter'". In: *Communications of the ACM* 66.12, pp. 35–39. DOI: 10.1145/3625833.

— (Oct. 25, 2024a). "Between the booms: AI in winter". In: *Communications of the ACM* 67.11, pp. 18–23. DOI: 10.1145/3688379.

— (Jan. 25, 2024b). "How the AI boom went bust". In: *Communications of the ACM* 67.2, pp. 22–26. DOI: 10.1145/3634901.

Hallinan, Christine Mary et al. (Feb. 1, 2024). "Seamless EMR data access: integrated governance, digital health and the OMOP-CDM". In: *BMJ Health & Care Informatics* 31.1, e100953. DOI: 10.1136/bmjhci-2023-100953.

Han, Yu, Aart C. Liefbroer & Cees H. Elzinga (Oct. 26, 2017). "Comparing methods of classifying life courses: sequence analysis and latent class analysis". In: *Longitudinal and Life Course Studies* 8.4 (4), pp. 319–341. DOI: 10.14301/llcs.v8i4.409.

Hanser, Michael, Claudio Di Ciccio & Jan Mendling (Jan. 27, 2016). "A novel framework for visualizing declarative process models". In: *Proceedings of the 8th ZEUS Workshop*. ZEUS 2016. Ed. by Christoph Hochreiner & Stefan Schulte. Vol. 1562. CEUR Workshop Proceedings. Vienna, AT: CEUR, pp. 5–12. URL: https://ceur-ws.org/Vol-1562/#paper1.

Harel, David (June 1, 1987). "Statecharts: a visual formalism for complex systems". In: *Science of Computer Programming* 8.3, pp. 231–274. DOI: 10.1016/0167-6423(87)90035-9.

HDRUK Phenotype Library (2023). Health Data Research UK. URL: https://phenotypes.healthdatagateway.org/ (retrieved 08/10/2023).

Heathfield, Heather (1999). "The rise and 'fall' of expert systems in medicine". In: *Expert Systems* 16.3, pp. 183–188. DOI: 10.1111/1468-0394.00107.

Hebbrecht, K. et al. (Dec. 23, 2020). "Understanding personalized dynamics to inform precision medicine: a dynamic time warp analysis of 255 depressed inpatients". In: *BMC Medicine* 18.1, p. 400. DOI: 10.1186/s12916-020-01867-5.

Hernán, Miguel A., David Clayton & Niels Keiding (June 1, 2011). "The Simpson's paradox unraveled". In: *International Journal of Epidemiology* 40.3, pp. 780–785. DOI: 10.1093/ije/dyr041.

Hicks, Michael Townsen, James Humphries & Joe Slater (June 8, 2024). "ChatGPT is bullshit". In: *Ethics and Information Technology* 26.2, p. 38. DOI: 10.1007/s10676-024-09775-5.

Hoare, C. A. R. (1985). *Communicating Sequential Processes*. Prentice Hall International. ISBN: 978-0-13-153289-2.

Hogan, Aidan et al. (July 2, 2021). "Knowledge graphs". In: *ACM Computing Surveys* 54.4, 71:1–71:37. DOI: 10.1145/3447772.

Hopcroft, John E. & Jeffrey D. Ullman (1979). *Introduction to Automata Theory, Languages, and Computation*. Addison-Wesley Series in Computer Science. Reading, MA, USA: Addison-Wesley. ISBN: 978-0-201-02988-8.

Hu, Jessica X. et al. (Feb. 15, 2019). "A large-cohort, longitudinal study determines precancer disease routes across different cancer types". In: *Cancer Research* 79.4, pp. 864–872. DOI: 10.1158/0008-5472.CAN-18-1677.

Hua-Gen Li, Michael et al. (Apr. 20, 2019). "Reliability of comorbidity scores derived from administrative data in the tertiary hospital intensive care setting: a cross-sectional study". In: *BMJ Health & Care Informatics* 26.1. DOI: 10.1136/bmjhci-2019-000016.

Huang, Zhengxing et al. (Jan. 2014). "Similarity measure between patient traces for clinical pathway analysis: problem, method, and applications". In: *IEEE Journal of Biomedical and Health Informatics* 18.1, pp. 4–14. DOI: 10.1109/JBHI.2013.2274281.

Hubert, Nicolas et al. (2024). "Do similar entities have similar embeddings?" In: *The Semantic Web*. ESWC 2024. Ed. by Albert Meroño Peñuela et al. Cham, CH: Springer Nature, pp. 3–21. DOI: 10.1007/978-3-031-60626-7_1.

Ilkou, Eleni & Maria Koutraki (Oct. 19, 2020). "Symbolic vs sub-symbolic AI methods: friends or enemies?" In: *Proceedings of the CIKM 2020 Workshops*. CIKM 2020. Ed. by Stefan Conrad & Ilaria Tiddi. Vol. 2699. CEUR Workshop Proceedings. Galway, IE: CEUR. URL: https://ceur-ws.org/Vol-2699/#paper06 (retrieved 08/18/2025).

Incitti, Francesca, Federico Urli & Lauro Snidaro (Jan. 1, 2023). "Beyond word embeddings: A survey". In: *Information Fusion* 89, pp. 418–436. DOI: 10.1016/j.inffus.2022.08.024.

Isern, David & Antonio Moreno (Dec. 1, 2008). "Computer-based execution of clinical guidelines: A review". In: *International Journal of Medical Informatics* 77.12, pp. 787–808. DOI: 10.1016/j.ijmedinf.2008.05.010.

ISO Technical Committee 215 (Dec. 2015). *Health Informatics — System of Concepts to Support Continuity of Care*. Geneva, CH. URL: https://www.iso.org/standard/58102.html (retrieved 12/15/2023).

Jain, Nitisha et al. (2021). "Do embeddings actually capture knowledge graph semantics?" In: *The Semantic Web*. ESWC 2021. Ed. by Ruben Verborgh et al. Cham, CH: Springer International, pp. 143–159. DOI: 10.1007/978-3-030-77385-4_9.

Jameson, J. Larry & Dan L. Longo (June 4, 2015). "Precision medicine — personalized, problematic, and promising". In: *New England Journal of Medicine* 372.23, pp. 2229–2234. DOI: 10.1056/NEJMsb1503104.

Jay, Caroline et al. (Feb. 2024). "Prioritize environmental sustainability in use of AI and data science methods". In: *Nature Geoscience* 17.2, pp. 106–108. DOI: 10.1038/s41561-023-01369-y.

Jeong, Young-Seon, Myong K. Jeong & Olufemi A. Omitaomu (Sept. 1, 2011). "Weighted dynamic time warping for time series classification". In: *Pattern Recognition* 44.9, pp. 2231–2240. DOI: 10.1016/j.patcog.2010.09.022.

Johnson, Alistair E.W. et al. (Jan. 6, 2023a). *MIMIC-IV*. Version 2.2. DOI: 10.13026/6MM1-EK67.

Johnson, Alistair E.W. et al. (Jan. 3, 2023b). "MIMIC-IV, a freely accessible electronic health record dataset". In: *Scientific Data* 10.1 (1), p. 1. DOI: 10.1038/s41597-022-01899-x.

Kaiser, Katharina & Mar Marcos (Feb. 10, 2016). "Leveraging workflow control patterns in the domain of clinical practice guidelines". In: *BMC Medical Informatics and Decision Making* 16.1, p. 20. DOI: 10.1186/s12911-016-0253-z.

Kane, Michael J. et al. (Dec. 17, 2023). "A compressed large language model embedding dataset of ICD 10 CM descriptions". In: *BMC Bioinformatics* 24.1, p. 482. DOI: 10.1186/s12859-023-05597-2.

Kapoor, Sayash & Arvind Narayanan (Sept. 8, 2023). "Leakage and the reproducibility crisis in machine-learning-based science". In: *Patterns* 4.9. DOI: 10.1016/j.patter.2023.100804.

Karpathakis, Kassandra, Jessica Morley & Luciano Floridi (Sept. 11, 2024). "A justifiable investment in AI for healthcare: aligning ambition with reality". In: *Minds and Machines* 34.4, p. 38. DOI: 10.1007/s11023-024-09692-y.

Kaul, Vivek, Sarah Enslin & Seth A. Gross (Oct. 1, 2020). "History of artificial intelligence in medicine". In: *Gastrointestinal Endoscopy* 92.4, pp. 807–812. DOI: 10.1016/j.gie.2020.06.040.

Kautz, Henry A. (2022). "The third AI summer: AAAI Robert S. Engelmore Memorial Lecture". In: *AI Magazine* 43.1, pp. 105–125. DOI: 10.1002/aaai.12036.

Kaymak, Uzay et al. (Oct. 2012). "On process mining in health care". In: *IEEE International Conference on Systems, Man, and Cybernetics*. SMC 2012, pp. 1859–1864. DOI: 10.1109/ICSMC.2012.6378009.

Khan, Arif, Shahadat Uddin & Uma Srinivasan (July 2018). "Comorbidity network for chronic disease: A novel approach to understand type 2 diabetes progression". In: *International Journal of Medical Informatics* 115, pp. 1–9. DOI: 10.1016/j.ijmedinf.2018.04.001.

Kho, Abel N. et al. (Mar. 2012). "Use of diverse electronic medical record systems to identify genetic risk for type 2 diabetes within a genome-wide association study". In: *Journal of the American Medical Informatics Association* 19.2, pp. 212–218. DOI: 10.1136/amiajnl-2011-000439.

Kilpatrick, Eric S. et al. (Jan. 17, 2022). "Controversies around the measurement of blood ketones to diagnose and manage diabetic ketoacidosis". In: *Diabetes Care* 45.2, pp. 267–272. DOI: 10.2337/dc21-2279.

Kinsman, Leigh et al. (May 27, 2010). "What is a clinical pathway? Development of a definition to inform the debate". In: *BMC Medicine* 8.1, p. 31. DOI: 10.1186/1741-7015-8-31.

Kirchner, Kathrin et al. (Jan. 1, 2023). "Patterns for modeling process variability in a healthcare context". In: *Business Process Management Journal* 30.1, pp. 1–27. DOI: 10.1108/BPMJ-10-2022-0500.

Klessascheck, Finn et al. (July 2, 2021). "Domain-specific event abstraction". In: *Business Information Systems*, pp. 117–126. DOI: 10.52825/bis.v1i.39.

Kuan, Valerie et al. (June 1, 2019). "A chronological map of 308 physical and mental health conditions from 4 million individuals in the English National Health Service". In: *The Lancet Digital Health* 1.2, e63–e77. DOI: 10.1016/S2589-7500(19)30012-3.

Kusuma, Guntur et al. (2020). "Process mining of disease trajectories: a feasibility study". In: *Proceedings of the 13th International Joint Conference*

on Biomedical Engineering Systems and Technologies — Vol. 5: HEALTHINF. BIOSTEC 2020. Vol. 5. Valetta, MT, pp. 705–712. DOI: 10.5220/0009166607050712.

Kusuma, Guntur et al. (May 27, 2021). "Process mining of disease trajectories: a literature review". In: *Studies in Health Technology and Informatics* 281, pp. 457–461. DOI: 10.3233/SHTI210200.

Lademann, Martin et al. (Sept. 2019). "Incorporating symptom data in longitudinal disease trajectories for more detailed patient stratification". In: *International Journal of Medical Informatics* 129, pp. 107–113. DOI: 10.1016/j.ijmedinf.2019.06.003.

Lamba, Ishan (June 2021). "Losing the numbers game: revisiting quality metrics through the spectrum of Goodhart's law". In: *European Journal of Emergency Medicine* 28.3, pp. 176–177. DOI: 10.1097/MEJ.0000000000000825.

Lang, Martin et al. (2008). "Process mining for clinical workflows: challenges and current limitations". In: *Proceedings of MIE2008 — The XXIst International Congress of the European Federation for Medical Informatics*. Ed. by Stig Kjær Andersen et al. Vol. 136. Studies in Health Technology and Informatics. Gothenburg, SE: IOS Press, pp. 229–234. URL: https://ebooks.iospress.nl/publication/11582.

Le, Quoc & Tomáš Mikolov (June 18, 2014). "Distributed representations of sentences and documents". In: *Proceedings of the 31st International Conference on Machine Learning*. ICML 2014. Vol. 32(2). Beijing, CN: PMLR, pp. 1188–1196. URL: https://proceedings.mlr.press/v32/le14.html (retrieved 03/12/2024).

Le Meur, Nolwenn, Fei Gao & Sahar Bayat (May 15, 2015). "Mining care trajectories using health administrative information systems: the use of state sequence analysis to assess disparities in prenatal care consumption". In: *BMC Health Services Research* 15, p. 200. DOI: 10.1186/s12913-015-0857-5.

Le Meur, Nolwenn et al. (June 2019). "Categorical state sequence analysis and regression tree to identify determinants of care trajectory in chronic disease: Example of end-stage renal disease". In: *Statistical Methods in Medical Research* 28.6, pp. 1731–1740. DOI: 10.1177/0962280218774811.

LeCun, Yann, Yoshua Bengio & Geoffrey Hinton (May 2015). "Deep learning". In: *Nature* 521.7553 (7553), pp. 436–444. DOI: 10.1038/nature14539.

Lee, Xing Ju et al. (Aug. 2019). "Review of methods and study designs of evaluations related to clinical pathways". In: *Australian Health Review* 43.4, pp. 448–456. DOI: 10.1071/AH17276.

Lehman, Eric et al. (June 13, 2023). "Do we still need clinical language models?" In: *Proceedings of the Conference on Health, Inference, and Learning*. CHIL

2023. Vol. 209. Proceedings of Machine Learning Research. Cambridge, MA, USA: PMLR, pp. 578–597. URL: https://proceedings.mlr.press/v209/eric23a (retrieved 10/22/2024).

Lejeune, Catherine et al. (June 1, 2010). "Socio-economic disparities in access to treatment and their impact on colorectal cancer survival". In: *International Journal of Epidemiology* 39.3, pp. 710–717. DOI: 10.1093/ije/dyq048.

Leonardi, G. et al. (2019). "Towards semantic process mining through knowledge-based trace abstraction". In: *Data-Driven Process Discovery and Analysis. SIMPDA 2017*. Ed. by Paolo Ceravolo, Maurice van Keulen & Kilian Stoffel. Lecture Notes in Business Information Processing. Cham, CH: Springer International, pp. 45–64. DOI: 10.1007/978-3-030-11638-5_3.

Levenshtein, Vladimir I. (Feb. 1966). "Binary codes capable of correcting deletions, insertions, and reversals". In: *Soviet Physics-Doklady* 10.8, pp. 707–710. ISSN: 0038-5689.

Li, Kenan et al. (Dec. 15, 2021). "Using dynamic time warping self-organizing maps to characterize diurnal patterns in environmental exposures". In: *Scientific Reports* 11.1, p. 24052. DOI: 10.1038/s41598-021-03515-1.

Lin, Yankai et al. (Jan. 25, 2015). "Learning entity and relation embeddings for knowledge graph completion". In: *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence*. AAAI 2015. Austin, TX, USA, pp. 2181–2187. ISBN: 978-0-262-51129-2.

Liyanage, Harshana et al. (Jan. 1, 2016). "Accessible Modelling of Complexity in Health and associated data flows: asthma as an exemplar". In: *Journal of Innovation in Health Informatics* 23.1. DOI: 10.14236/jhi.v23i1.863.

Lomotan, E. A. et al. (Dec. 1, 2010). "How "should" we write guideline recommendations? Interpretation of deontic terminology in clinical practice guidelines: survey of the health services community". In: *BMJ Quality & Safety* 19.6, pp. 509–513. DOI: 10.1136/qshc.2009.032565.

López, Hugo A. & Vít Dexter Simon (Nov. 23, 2022). "How to (re)design declarative process notations? A view from the lens of cognitive effectiveness frameworks". In: *Proceedings of the Forum at Practice of Enterprise Modeling 2022*. PoEM 2022. Ed. by Tony Clark et al. Vol. 3327. CEUR Workshop Proceedings. London, UK: CEUR, pp. 81–97. URL: https://ceur-ws.org/Vol-3327/#paper08.

Loveday, Chey et al. (June 1, 2021). "Prioritisation by FIT to mitigate the impact of delays in the 2-week wait colorectal cancer referral pathway during the COVID-19 pandemic: a UK modelling study". In: *Gut* 70.6, pp. 1053–1060. DOI: 10.1136/gutjnl-2020-321650.

Lucivero, Federica (June 21, 2024). "AI and environmental sustainability". In: *Handbook on Public Policy and Artificial Intelligence*. Ed. by Regine Paul, Emma Carmel & Jennifer Cobbe. Edward Elgar Publishing, pp. 158–169. ISBN: 978-1-80392-217-1.

Luzi, Daniela et al. (June 1, 2019). "Modelling collaboration of primary and secondary care for children with complex care needs: long-term ventilation as an example". In: *European Journal of Pediatrics* 178.6, pp. 891–901. DOI: 10.1007/s00431-019-03367-y.

Madandola, Olatunde O. et al. (Jan. 1, 2024). "The relationship between electronic health records user interface features and data quality of patient clinical information: an integrative review". In: *Journal of the American Medical Informatics Association* 31.1, pp. 240–255. DOI: 10.1093/jamia/ocad188.

Maier, David (Apr. 1, 1978). "The complexity of some problems on subsequences and supersequences". In: *Journal of the ACM* 25.2, pp. 322–336. DOI: 10.1145/322063.322075.

Makadia, Rupa et al. (Dec. 1, 2023). "Evaluating the impact of alternative phenotype definitions on incidence rates across a global data network". In: *JAMIA Open* 6.4, ooad096. DOI: 10.1093/jamiaopen/ooad096.

Mans, Ronny et al. (2008). "Process mining techniques: an application to stroke care". In: *Proceedings of MIE2008 — The XXIst International Congress of the European Federation for Medical Informatics*. Ed. by Stig Kjær Andersen et al. Vol. 136. Studies in Health Technology and Informatics. Gothenburg, SE: IOS Press, pp. 573–578. ISBN: 978-1-60750-333-0.

Marijnen, C.A.M. et al. (July 1, 2020). "International expert consensus statement regarding radiotherapy treatment options for rectal cancer during the COVID 19 pandemic". In: *Radiotherapy and Oncology* 148, pp. 213–215. DOI: 10.1016/j.radonc.2020.03.039.

Marin-Castro, Heidy M. & Edgar Tello-Leal (Jan. 2021). "Event log preprocessing for process mining: a review". In: *Applied Sciences* 11.22, p. 10556. DOI: 10.3390/app112210556.

Martin, Niels, Isabeau Gielen & Jochen Bergs (Jan. 3, 2024). "Process mining using electronic health records data – quo vadis? Reflections from observing nurses' activities and data registration behavior". In: *Proceedings of the 57th Hawaii International Conference on System Sciences*. HICSS 2023. Honolulu, HI, USA: University of Hawai'i at Mānoa, pp. 3707–3716. ISBN: 978-0-9981331-7-1.

Mattson, Christopher, Reamer L. Bushardt & Anthony R. Artino (Feb. 2021). ""When a measure becomes a target, it ceases to be a good measure"". In:

Journal of Graduate Medical Education 13.1, pp. 2–5. DOI: 10.4300/JGME-D-20-01492.1.

Mayer, Claus-Dieter, Julie Lorent & Graham W. Horgan (Mar. 2, 2011). "Exploratory analysis of multiple omics datasets using the adjusted RV coefficient". In: *Statistical Applications in Genetics and Molecular Biology* 10.1. DOI: 10.2202/1544-6115.1540.

McKee, Martin (May 16, 2016). "The weekend effect: now you see it, now you don't". In: *BMJ* 353, p. i2750. DOI: 10.1136/bmj.i2750.

Meacock, Rachel et al. (Jan. 2017). "Higher mortality rates amongst emergency patients admitted to hospital at weekends reflect a lower probability of admission". In: *Journal of Health Services Research & Policy* 22.1, pp. 12–19. DOI: 10.1177/1355819616649630.

Meert, Wannes et al. (Aug. 11, 2020). *DTAIdistance*. Version 2.3.10. DOI: 10.5281/zenodo.5901139.

Mesbah, R. et al. (2024). "Dynamic time warp analysis of individual symptom trajectories in individuals with bipolar disorder". In: *Bipolar Disorders* 26.1, pp. 44–57. DOI: 10.1111/bdi.13340.

Middlehurst, Matthew, Patrick Schäfer & Anthony Bagnall (July 1, 2024). "Bake off redux: a review and experimental evaluation of recent time series classification algorithms". In: *Data Mining and Knowledge Discovery* 38.4, pp. 1958–2031. DOI: 10.1007/s10618-024-01022-1.

Mikolai, Julia & Mark Lyons-Amos (Apr. 27, 2017). "Longitudinal methods for life course research: A comparison of sequence analysis, latent class growth models, and multi-state event history models for studying partnership transitions". In: *Longitudinal and Life Course Studies* 8.2 (2), pp. 191–208. DOI: 10.14301/llcs.v8i2.415.

Mikolov, Tomáš et al. (Dec. 5, 2013a). "Distributed representations of words and phrases and their compositionality". In: *Proceedings of the 26th International Conference on Neural Information Processing Systems - Volume 2*. NeurIPS 2013. Red Hook, NY, USA: Curran Associates Inc., pp. 3111–3119. ISBN: 978-1-63266-024-4.

Mikolov, Tomáš et al. (May 2013b). "Efficient estimation of word representations in vector space". In: *1st International Conference on Learning Representations (Workshop Track Proceedings)*. ICLR 2013. Ed. by Yoshua Bengio & Yann LeCun. Scottsdale, AZ, USA. DOI: 10.48550/arXiv.1301.3781.

Ministry of Housing, Communities & Local Government (2019). *English indices of deprivation 2019*. Ministry of Housing, Communities & Local Government. URL: https://www.gov.uk/government/statistics/english-indices-of-deprivation-2019 (retrieved 04/22/2024).

Mitchell, Jeff & Mirella Lapata (June 2008). "Vector-based models of semantic composition". In: *Proceedings of ACL-08: HLT*. ACL-HLT 2008. Ed. by Johanna D. Moore et al. Columbus, OH, USA: Association for Computational Linguistics, pp. 236–244. URL: https://aclanthology.org/P08-1028 (retrieved 04/09/2024).

Møller, Henrik et al. (Jan. 1, 2012). "Colorectal cancer survival in socioeconomic groups in England: variation is mainly in the short term after diagnosis". In: *European Journal of Cancer* 48.1, pp. 46–53. DOI: 10.1016/j.ejca.2011.05.018.

Moran, Brendan et al. (2017). "Association of Coloproctology of Great Britain & Ireland (ACPGBI): Guidelines for the Management of Cancer of the Colon, Rectum and Anus (2017) – Surgical Management". In: *Colorectal Disease* 19.S1, pp. 18–36. DOI: 10.1111/codi.13704.

Morley, Jessica & Joe Zhang (Nov. 29, 2023). "A controversial new federated data platform for the NHS in England". In: *BMJ* 383, p2776. DOI: 10.1136/bmj.p2776.

Morley, Katherine I. et al. (Nov. 4, 2014). "Defining disease phenotypes using national linked electronic health records: a case study of atrial fibrillation". In: *PLOS ONE* 9.11, e110900. DOI: 10.1371/journal.pone.0110900.

Morris, Eva J. A. et al. (Dec. 1, 2008). "Unacceptable variation in abdominoperineal excision rates for rectal cancer: time to intervene?" In: *Gut* 57.12, pp. 1690–1697. DOI: 10.1136/gut.2007.137877.

Morris, Eva J. A. et al. (Aug. 2016). "Wide variation in the use of radiotherapy in the management of surgically treated rectal cancer across the English National Health Service". In: *Clinical Oncology* 28.8, pp. 522–531. DOI: 10.1016/j.clon.2016.02.002.

Morris, Eva J. A. et al. (Mar. 1, 2021). "Impact of the COVID-19 pandemic on the detection and management of colorectal cancer in England: a population-based study". In: *The Lancet Gastroenterology & Hepatology* 6.3, pp. 199–208. DOI: 10.1016/S2468-1253(21)00005-4.

Morris, Melanie et al. (Sept. 1, 2020). "Understanding the link between health systems and cancer survival: A novel methodological approach using a system-level conceptual model". In: *Journal of Cancer Policy* 25, p. 100233. DOI: 10.1016/j.jcpo.2020.100233.

Mortensen, Jonathan M et al. (Mar. 1, 2015). "Using the wisdom of the crowds to find critical errors in biomedical ontologies: a study of SNOMED CT". In: *Journal of the American Medical Informatics Association* 22.3, pp. 640–648. DOI: 10.1136/amiajnl-2014-002901.

Mueen, Abdullah & Eamonn Keogh (Aug. 13, 2016). "Extracting optimal performance from dynamic time warping". In: *Proceedings of the 22nd*

ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. KDD 2016. New York, NY, USA: Association for Computing Machinery, pp. 2129–2130. DOI: 10.1145/2939672.2945383.

Muller, Patrick & Laura Woods (Aug. 1, 2022). "Multiple imputation to minimise bias from missing stage information in estimates of early cancer diagnosis in England: a population-based study". In: *Cancer Epidemiology* 79, p. 102198. DOI: 10.1016/j.canep.2022.102198.

Mulyar, Nataliya, Wil M. P. van der Aalst & Mor Peleg (Nov. 1, 2007). "A pattern-based analysis of clinical computer-interpretable guideline modeling languages". In: *Journal of the American Medical Informatics Association* 14.6, pp. 781–787. DOI: 10.1197/jamia.M2389.

Munoz-Gama, Jorge et al. (Mar. 1, 2022). "Process mining for healthcare: characteristics and challenges". In: *Journal of Biomedical Informatics* 127, p. 103994. DOI: 10.1016/j.jbi.2022.103994.

Nab, Linda et al. (2024). "OpenSAFELY: A platform for analysing electronic health records designed for reproducible research". In: *Pharmacoepidemiology and Drug Safety* 33.6, e5815. DOI: 10.1002/pds.5815.

Nalisnick, Eric et al. (Apr. 11, 2016). "Improving document ranking with dual word embeddings". In: *Proceedings of the 25th International Conference Companion on World Wide Web*. WWW 2016. Geneva, CH, pp. 83–84. DOI: 10.1145/2872518.2889361.

National Disease Registration Service (n.d.). *Linking Treatment Tables: Chemotherapy, Tumour Resections, and Radiotherapy (Cancer Analysis System Standard Operating Procedure)*. ver 4.9. URL: https://digital.nhs.uk/binaries/content/assets/website-assets/national-disease-registration-service/cancer-data-training-materials/linking-treatment-data-for-cancer-diagnoses-cas-sop-v4.9.html.

NDRS (2021). *Cancer treatments*. National Disease Registration Service. URL: https://digital.nhs.uk/ndrs/data/data-outputs/cancer-data-hub/cancer-treatments (retrieved 08/06/2024).

Neal, Richard D. et al. (Mar. 1, 2007). "Stage, survival and delays in lung, colorectal, prostate and ovarian cancer: comparison between diagnostic routes". In: *The British Journal of General Practice* 57.536, pp. 212–219. ISSN: 0960-1643.

Neal, Richard D. et al. (Feb. 2014). "Comparison of cancer diagnostic intervals before and after implementation of NICE guidelines: analysis of data from the UK General Practice Research Database". In: *British Journal of Cancer* 110.3, pp. 584–592. DOI: 10.1038/bjc.2013.791.

Neal, Richard D. et al. (Mar. 31, 2015). "Is increased time to diagnosis and treatment in symptomatic cancer associated with poorer outcomes? Sys-

tematic review". In: *British Journal of Cancer* 112 (Suppl 1), S92–S107. DOI: 10.1038/bjc.2015.48.

Nguyen-Duc, Thanh et al. (2021). "Deep EHR spotlight: a framework and mechanism to highlight events in electronic health records for explainable predictions". In: *AMIA Joint Summits on Translational Science Proceedings*. AMIA Informatics Summit 2021. Vol. 2021. Online, pp. 475–484. PMID: 34457163.

NHS Constitution for England (Mar. 8, 2012). Department for Health and Social Care. URL: https://www.gov.uk/government/publications/the-nhs-constitution-for-england (retrieved 08/06/2024).

NHS England (2020). *Information standard SCCI0034: SNOMED CT*. URL: https://digital.nhs.uk/data-and-information/information-standards/information-standards-and-data-collections-including-extractions/publications-and-notifications/standards-and-collections/scci0034-snomed-ct (retrieved 10/17/2024).

— (Aug. 15, 2023a). *Cancer Waiting Times Review — Models of care and measurement: consultation response*. URL: https://www.england.nhs.uk/wp-content/uploads/2023/08/PRN00654i-cancer-waiting-times-review-consultation-response.pdf (retrieved 09/16/2024).

— (Apr. 2023b). *Implementing a Timed Colorectal Cancer Diagnostic Pathway: Guidance for Local Health and Care Systems*. ver. 3.2. URL: https://www.england.nhs.uk/wp-content/uploads/2018/04/B2119-implementing-timed-colorectal-cancer-diagnostic-pathway-2.pdf (retrieved 09/16/2024).

— (Apr. 2023c). *National Clinical Coding Standards ICD-10*. 5th ed., ver. 9.0. URL: https://classbrowser.nhs.uk/ref_books/ICD-10_2023_5th_Ed_NCCS.pdf.

— (Apr. 2024). *National Clinical Coding Standards OPCS-4*. ver. 11.1. URL: https://classbrowser.nhs.uk/ref_books/OPCS-4.10_NCCS-2024.pdf.

NHS Wales (Nov. 2023). *National Optimal Pathway for Colorectal Cancer*. 2nd ed. URL: https://executive.nhs.wales/functions/networks-and-planning/cancer/wcn-documents/clinician-hub/csg-pathways-and-associated-documents/colorectal-nop/ (retrieved 09/16/2024).

NICE (June 23, 2015). *Suspected cancer: recognition and referral*. National Institute for Health and Care Excellence. URL: https://www.nice.org.uk/guidance/ng12 (retrieved 08/22/2024).

— (Jan. 29, 2020). *Colorectal cancer*. National Institute for Health and Care Excellence. URL: https://www.nice.org.uk/guidance/ng151 (retrieved 08/22/2024).

NICE (Sept. 27, 2023a). *Artificial intelligence technologies to aid contouring for radiotherapy treatment planning: early value assessment*. National Institute for Health and Care Excellence. URL: https://www.nice.org.uk/guidance/hte11 (retrieved 08/29/2024).

— (Sept. 28, 2023b). *Artificial intelligence-derived software to analyse chest X-rays for suspected lung cancer in primary care referrals: early value assessment*. National Institute for Health and Care Excellence. URL: https://www.nice.org.uk/guidance/hte12 (retrieved 08/29/2024).

— (Aug. 24, 2023c). *Quantitative faecal immunochemical testing to guide colorectal cancer pathway referral in primary care*. National Institute for Health and Care Excellence. URL: https://www.nice.org.uk/guidance/dg56 (retrieved 08/22/2024).

Nicholson, Brian D. et al. (Jan. 1, 2018). "The Suspected CANcer (SCAN) pathway: protocol for evaluating a new standard of care for patients with non-specific symptoms of cancer". In: *BMJ Open* 8.1, e018168. DOI: 10.1136/bmjopen-2017-018168.

Nickel, Maximilian, Volker Tresp & Hans-Peter Kriegel (June 28, 2011). "A three-way model for collective learning on multi-relational data". In: *Proceedings of the 28th International Conference on International Conference on Machine Learning*. ICML 2011. Madison, WI, USA: Omnipress, pp. 809–816. ISBN: 978-1-4503-0619-5.

— (Apr. 16, 2012). "Factorizing YAGO: scalable machine learning for linked data". In: *Proceedings of the 21st International Conference on World Wide Web*. WWW 2012. New York, NY, USA: Association for Computing Machinery, pp. 271–280. DOI: 10.1145/2187836.2187874.

Nouraei, Seyed Ahmad Reza et al. (June 1, 2016). "Accuracy of clinician-clinical coder information handover following acute medical admissions: implication for using administrative datasets in clinical outcomes management". In: *Journal of Public Health* 38.2, pp. 352–362. DOI: 10.1093/pubmed/fdv041.

O'Malley, Kimberly J et al. (Oct. 2005). "Measuring diagnoses: ICD code accuracy". In: *Health Services Research* 40 (5 Part 2), pp. 1620–1639. DOI: 10.1111/j.1475-6773.2005.00444.x.

Oliart, Eimy, Eric Rojas & Daniel Capurro (June 1, 2022). "Are we ready for conformance checking in healthcare? Measuring adherence to clinical guidelines: A scoping systematic literature review". In: *Journal of Biomedical Informatics* 130, p. 104076. DOI: 10.1016/j.jbi.2022.104076.

Oliveira, Tiago, Paulo Novais & José Neves (Dec. 2014). "Development and implementation of clinical guidelines: An artificial intelligence perspective".

In: *Artificial Intelligence Review* 42.4, pp. 999–1027. DOI: 10.1007/s10462-013-9402-2.

OMG (Dec. 2013). *Business Process Model and Notation*. ver. 2.0.2. Object Management Group. URL: http://www.omg.org/spec/BPMN (retrieved 08/02/2022).

— (Dec. 2017). *OMG Unified Modeling Language*. ver. 2.5.1. Object Management Group. URL: https://www.omg.org/spec/UML/ (retrieved 08/23/2024).

OpenCodelists (2023). Bennett Institute for Applied Data Science, University of Oxford. URL: https://www.opencodelists.org/ (retrieved 08/10/2023).

Oughtibridge, Nicholas (Oct. 2019). *A system of concepts for the continuity of care: a visualisation of a model conforming to International Standard EN ISO 13940:2016*. URL: https://contsys.org/ (retrieved 12/15/2023).

Owens, Brian (Sept. 18, 2024). "Rage against machine learning driven by profit". In: *Nature* 633.8030, S6–S9. DOI: 10.1038/d41586-024-02985-3.

Pai, Shraddha & Gary D. Bader (Sept. 14, 2018). "Patient similarity networks for precision medicine". In: *Journal of Molecular Biology* 430 (18 Part A), pp. 2924–2938. DOI: 10.1016/j.jmb.2018.05.037.

Pai, Shraddha et al. (Mar. 2019). "netDx: interpretable patient classification using integrated patient similarity networks". In: *Molecular Systems Biology* 15.3, e8497. DOI: 10.15252/msb.20188497.

Paik, Hyojung et al. (Oct. 15, 2019). "Tracing diagnosis trajectories over millions of patients reveal an unexpected risk in schizophrenia". In: *Scientific Data* 6.1 (1), p. 201. DOI: 10.1038/s41597-019-0220-5.

Panteli, Dimitra et al. (2019). "Clinical practice guidelines as a quality strategy". In: *Improving Healthcare Quality in Europe: Characteristics, Effectiveness and Implementation of Different Strategies*. Copenhagen, DK: European Observatory on Health Systems and Policies. ISBN: 978 92 648 0590 3.

Papez, Vaclav et al. (July 1, 2021). "Transforming and evaluating electronic health record disease phenotyping algorithms using the OMOP common data model: a case study in heart failure". In: *JAMIA Open* 4.3, ooab001. DOI: 10.1093/jamiaopen/ooab001.

Papez, Vaclav et al. (Jan. 1, 2023). "Transforming and evaluating the UK Biobank to the OMOP Common Data Model for COVID-19 research and beyond". In: *Journal of the American Medical Informatics Association* 30.1, pp. 103–111. DOI: 10.1093/jamia/ocac203.

Paulheim, Heiko & Christian Bizer (2013). "Type inference on noisy RDF data". In: *The Semantic Web*. ISWC 2013. Ed. by Harith Alani et al. Berlin

& Heidelberg, DE: Springer, pp. 510–525. DOI: 10.1007/978-3-642-41335-3_32.

Pecoraro, Fabrizio & Daniela Luzi (Jan. 2022). "Using Unified Modeling Language to analyze business processes in the delivery of child health services". In: *International Journal of Environmental Research and Public Health* 19.20, p. 13456. DOI: 10.3390/ijerph192013456.

Peleg, Mor (Aug. 1, 2013). "Computer-interpretable clinical guidelines: A methodological review". In: *Journal of Biomedical Informatics* 46.4, pp. 744–763. DOI: 10.1016/j.jbi.2013.06.009.

Perer, Adam, Fei Wang & Jianying Hu (Aug. 1, 2015). "Mining and exploring care pathways from electronic medical records with visual analytics". In: *Journal of Biomedical Informatics* 56, pp. 369–378. DOI: 10.1016/j.jbi.2015.06.020.

Pine, Kathleen H. & Max Liboiron (Apr. 18, 2015). "The politics of measurement and action". In: *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*. CHI 2015. New York, NY, USA: Association for Computing Machinery, pp. 3147–3156. DOI: 10.1145/2702123.2702298.

PlantUML (2023). *Drawing UML with PlantUML: PlantUML Language Reference Guide*. PlantUML. URL: https://plantuml.com/guide (retrieved 11/20/2024).

Pokharel, Suresh et al. (Aug. 1, 2020). "Temporal tree representation for similarity computation between medical patients". In: *Artificial Intelligence in Medicine* 108, p. 101900. DOI: 10.1016/j.artmed.2020.101900.

Portisch, Jan, Nicolas Heist & Heiko Paulheim (Jan. 1, 2022). "Knowledge graph embedding for data mining vs. knowledge graph embedding for link prediction – two sides of the same coin?" In: *Semantic Web* 13.3, pp. 399–422. DOI: 10.3233/SW-212892.

Pufahl, Luise et al. (July 1, 2022). "BPMN in healthcare: challenges and best practices". In: *Information Systems* 107, p. 102013. DOI: 10.1016/j.is.2022.102013.

Ram, A et al. (Feb. 1, 2022). "Transphobia, encoded: an examination of transspecific terminology in SNOMED CT and ICD-10-CM". In: *Journal of the American Medical Informatics Association* 29.2, pp. 404–410. DOI: 10.1093/jamia/ocab200.

Rastogi, Nidhi & Mohammed J. Zaki (May 7, 2020). "Personal health knowledge graphs for patients". In: *Workshop on Personal Health Knowledge Graphs*. KGC 2020. Online. DOI: 10.48550/arXiv.2004.00071.

Rebuge, Álvaro & Diogo R. Ferreira (Apr. 1, 2012). "Business process analysis in healthcare environments: a methodology based on process mining". In: *Information Systems* 37.2, pp. 99–116. DOI: 10.1016/j.is.2011.01.003.

Reich, Christian et al. (Jan. 4, 2024). "OHDSI Standardized Vocabularies—a large-scale centralized reference ontology for international data harmonization". In: *Journal of the American Medical Informatics Association*, ocad247. DOI: 10.1093/jamia/ocad247.

Remy, Simon et al. (2020). "Event log generation in a health system: a case study". In: *Business Process Management*. BPM 2020. Ed. by Dirk Fahland et al. Vol. 12168. Lecture Notes in Computer Science. Cham, CH: Springer International, pp. 505–522. DOI: 10.1007/978-3-030-58666-9_29.

Ren, Ruohan, Changchuan Yin & Stephen S.-T. Yau (Sept. 2022). "Kmer2vec: A Novel Method for Comparing DNA Sequences by word2vec Embedding". In: *Journal of Computational Biology* 29.9, pp. 1001–1021. DOI: 10.1089/cmb.2021.0536.

Ristoski, Petar & Heiko Paulheim (2016). "RDF2Vec: RDF Graph Embeddings for Data Mining". In: *The Semantic Web*. ISWC 2016. Ed. by Paul Groth et al. Cham, CH: Springer International, pp. 498–514. DOI: 10.1007/978-3-319-46523-4_30.

Rivault, Yann, Nolwenn Le Meur & Olivier Dameron (2017). "A similarity measure based on care trajectories as sequences of sets". In: *Artificial Intelligence in Medicine*. Ed. by Annette ten Teije et al. Lecture Notes in Computer Science. Cham, CH: Springer International, pp. 278–282. DOI: 10.1007/978-3-319-59758-4_32.

Robert, P. & Y. Escoufier (1976). "A unifying tool for linear multivariate statistical methods: the RV-coefficient". In: *Journal of the Royal Statistical Society, Series C (Applied Statistics)* 25.3, pp. 257–265. DOI: 10.2307/2347233.

Roberts, Michael et al. (Mar. 2021). "Common pitfalls and recommendations for using machine learning to detect and prognosticate for COVID-19 using chest radiographs and CT scans". In: *Nature Machine Intelligence* 3.3, pp. 199–217. DOI: 10.1038/s42256-021-00307-0.

Rojas, Eric et al. (June 1, 2016). "Process mining in healthcare: a literature review". In: *Journal of Biomedical Informatics* 61, pp. 224–236. DOI: 10.1016/j.jbi.2016.04.007.

Roth, Leonard et al. (Dec. 26, 2022). "Identifying common patterns of health services use: a longitudinal study of older Swiss adults' care trajectories". In: *BMC Health Services Research* 22.1, p. 1586. DOI: 10.1186/s12913-022-08987-z.

Rotter, Thomas et al. (Mar. 17, 2010). "Clinical pathways: effects on professional practice, patient outcomes, length of stay and hospital costs". In: *The Cochrane Database of Systematic Reviews* 3, p. CD006632. DOI: 10.1002/14651858.CD006632.pub2.

Rotter, Thomas et al. (June 18, 2012). "The quality of the evidence base for clinical pathway effectiveness: room for improvement in the design of evaluation trials". In: *BMC Medical Research Methodology* 12, p. 80. DOI: 10.1186/1471-2288-12-80.

Rotter, Thomas et al. (2019). "Clinical pathways as a quality strategy". In: *Improving Healthcare Quality in Europe: Characteristics, Effectiveness and Implementation of Different Strategies*. European Observatory on Health Systems and Policies. ISBN: 978 92 648 0590 3.

Roux, Jonathan, Olivier Grimaud & Emmanuelle Leray (June 2019). "Use of state sequence analysis for care pathway analysis: The example of multiple sclerosis". In: *Statistical Methods in Medical Research* 28.6, pp. 1651–1663. DOI: 10.1177/0962280218772068.

Royal College of Radiologists (2019). *The Timely Delivery of Radical Radiotherapy: Guidelines for the Management of Unscheduled Treatment Interruptions*. 4th ed. London, UK: Royal College of Radiologists. URL: https://www.rcr.ac.uk/media/z5jgmrhd/rcr-publications_the-timely-delivery-of-radical-radiotherapy-guidelines-for-the-management-of-unscheduled-treatment-interruptions-4th-edition_january-2019.pdf.

Rücker, Gerta & Martin Schumacher (May 30, 2008). "Simpson's paradox visualized: the example of the Rosiglitazone meta-analysis". In: *BMC Medical Research Methodology* 8.1, p. 34. DOI: 10.1186/1471-2288-8-34.

Ruffinelli, Daniel, Samuel Broscheit & Rainer Gemulla (Sept. 25, 2019). "You CAN teach an old dog new tricks! On training knowledge graph embeddings". In: *International Conference on Learning Representations*. ICLR 2020. Online. URL: https://openreview.net/forum?id=BkxSmlBFvr (retrieved 11/05/2024).

Rumbaugh, James, Ivar Jacobson & Grady Booch (1999). *The Unified Modeling Language Reference Manual*. Reading, MA, USA: Addison-Wesley. ISBN: 0-201-30998-X.

Russell, Nick et al. (Jan. 1, 2006). "On the suitability of UML 2.0 activity diagrams for business process modelling". In: *Proceedings of the 3rd Asia-Pacific Conference on Conceptual Modelling*. APCCM 2006. Vol. 53. Hobart, AU, pp. 95–104. ISBN: 978-1-920682-35-4.

Russell, Stuart & Peter Norvig (2021). *Artificial Intelligence: A Modern Approach*. 4th (Global). Harlow, UK: Pearson Education. ISBN: 978-1-292-40117-1.

Sambasivan, Nithya et al. (May 7, 2021). ""Everyone wants to do the model work, not the data work": data cascades in high-stakes AI". In: *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. CHI 2021. New York, NY, USA: Association for Computing Machinery, pp. 1–15. DOI: 10.1145/3411764.3445518.

Savaré, Laura et al. (July 29, 2023). "Capturing the variety of clinical pathways in patients with schizophrenic disorders through state sequences analysis". In: *BMC Medical Research Methodology* 23.1, p. 174. DOI: 10.1186/s12874-023-01993-7.

Savino, Mariachiara et al. (May 17, 2023). "A process mining approach for clinical guidelines compliance: real-world application in rectal cancer". In: *Frontiers in Oncology* 13. DOI: 10.3389/fonc.2023.1090076.

Scheuerlein, Hubert et al. (June 1, 2012). "New methods for clinical pathways—Business Process Modeling Notation (BPMN) and Tangible Business Process Modeling (t.BPM)". In: *Langenbeck's Archives of Surgery* 397.5, pp. 755–761. DOI: 10.1007/s00423-012-0914-z.

Schrodt, Jens et al. (Mar. 12, 2020). "Graph-representation of patient data: a systematic literature review". In: *Journal of Medical Systems* 44.4, p. 86. DOI: 10.1007/s10916-020-1538-4.

Scott, Philip et al. (2023). "Modelling clinical narrative as computable knowledge: the NICE computable implementation guidance project". In: *Learning Health Systems* 7.4, e10394. DOI: 10.1002/lrh2.10394.

Shojaee, Parshin et al. (July 18, 2025). *The illusion of thinking: understanding the strengths and limitations of reasoning models via the lens of problem complexity*. DOI: 10.48550/arXiv.2506.06941. arXiv: 2506.06941 [cs]. Pre-published.

Shortliffe, Edward H. (Dec. 1986). "Medical expert systems—knowledge tools for physicians". In: *Western Journal of Medicine* 145.6, pp. 830–839. PMID: 3811349.

Shrestha, Anne et al. (2019). "Quality of life versus length of life considerations in cancer patients: a systematic literature review". In: *Psycho-Oncology* 28.7, pp. 1367–1380. DOI: 10.1002/pon.5054.

Shumailov, Ilia et al. (July 2024). "AI models collapse when trained on recursively generated data". In: *Nature* 631.8022, pp. 755–759. DOI: 10.1038/s41586-024-07566-y.

Siggaard, Troels et al. (Oct. 2, 2020). "Disease trajectory browser for exploring temporal, population-wide disease progression patterns in 7.2 million

Danish patients". In: *Nature Communications* 11.1 (1), p. 4952. DOI: 10.1038/s41467-020-18682-4.

Simpson, E.H. (1951). "The interpretation of interaction in contingency tables". In: *Journal of the Royal Statistical Society, Series B (Statistical Methodology)* 13.2, pp. 238–241. ISSN: 0035-9246.

Singhal, Amit (May 16, 2012). *Introducing the Knowledge Graph: things, not strings*. Google. URL: https://blog.google/products/search/introducing-knowledge-graph-things-not/ (retrieved 03/28/2022).

Skelac, Ines & Andrej Jandrić (2020). "Meaning as use: from Wittgenstein to Google's Word2vec". In: *Guide to Deep Learning Basics: Logical, Historical and Philosophical Perspectives*. Ed. by Sandro Skansi. Cham, CH: Springer, pp. 41–53. DOI: 10.1007/978-3-030-37591-1_5.

Slater, Luke T., Georgios V. Gkoutos & Robert Hoehndorf (Dec. 15, 2020). "Towards semantic interoperability: finding and repairing hidden contradictions in biomedical ontologies". In: *BMC Medical Informatics and Decision Making* 20.10, p. 311. DOI: 10.1186/s12911-020-01336-2.

SNOMED International (Aug. 24, 2022). *Expression Constraint Language - Specification and Guide*. ver 2.1. International Health Terminology Standards Development Organisation. URL: http://snomed.org/ecl (retrieved 04/12/2023).

Soukup, Tayana et al. (2020). "A multicentre cross-sectional observational study of cancer multidisciplinary teams: analysis of team decision making". In: *Cancer Medicine* 9.19, pp. 7083–7099. DOI: 10.1002/cam4.3366.

Stothers, Jessica A.M. & Andrew Nguyen (2020). "Can Neo4j replace PostgreSQL in healthcare?" In: *AMIA Joint Summits on Translational Science Proceedings*. AMIA Informatics Summit 2020. Vol. 2020. Online, pp. 646–653. PMID: 32477687.

Strubell, Emma, Ananya Ganesh & Andrew McCallum (July 2019). "Energy and policy considerations for deep learning in NLP". In: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. ACL 2019. Florence, IT: Association for Computational Linguistics, pp. 3645–3650. DOI: 10.18653/v1/P19-1355.

Studer, Matthias & Gilbert Ritschard (Feb. 1, 2016). "What matters in differences between life trajectories: a comparative review of sequence dissimilarity measures". In: *Journal of the Royal Statistical Society, Series A (Statistics in Society)* 179.2, pp. 481–511. DOI: 10.1111/rssa.12125.

Sudlow, Cathie (Nov. 8, 2024). *Uniting the UK's Health Data: A Huge Opportunity for Society*. DOI: 10.5281/zenodo.13353747.

Sun, Zhiqing et al. (May 7, 2019). "RotatE: knowledge graph embedding by relational rotation in complex space". In: *7th International Conference*

on Learning Representations. ICLR 2019. New Orleans, LA, USA. DOI: 10.48550/arXiv.1902.10197.

Sung, Hyuna et al. (Dec. 11, 2024). "Colorectal cancer incidence trends in younger versus older adults: an analysis of population-based cancer registry data". In: *The Lancet Oncology*. DOI: 10.1016/S1470-2045(24)00600-4.

Sutton, Reed T. et al. (Feb. 6, 2020). "An overview of clinical decision support systems: benefits, risks, and strategies for success". In: *npj Digital Medicine* 3.1, p. 17. DOI: 10.1038/s41746-020-0221-y.

Syriopoulou, Elisavet et al. (May 2019). "Understanding the impact of socioeconomic differences in colorectal cancer survival: potential gain in life-years". In: *British Journal of Cancer* 120.11, pp. 1052–1058. DOI: 10.1038/s41416-019-0455-0.

Tamm, Andres et al. (June 2022). "Establishing a colorectal cancer research database from routinely collected health data: the process and potential from a pilot study". In: *BMJ Health & Care Informatics* 29.1, e100535. DOI: 10.1136/bmjhci-2021-100535.

Tax, Niek, Natalia Sidorova & Wil M. P. van der Aalst (Feb. 1, 2019). "Discovering more precise process models from event logs by filtering out chaotic activities". In: *Journal of Intelligent Information Systems* 52.1, pp. 107–139. DOI: 10.1007/s10844-018-0507-6.

Taylor, John C. et al. (2021). "Addressing the variation in adjuvant chemotherapy treatment for colorectal cancer: Can a regional intervention promote national change?" In: *International Journal of Cancer* 148.4, pp. 845–856. DOI: 10.1002/ijc.33261.

Thygesen, Johan H. et al. (July 1, 2022). "COVID-19 trajectories among 57 million adults in England: a cohort study using electronic health records". In: *The Lancet Digital Health* 4.7, e542–e557. DOI: 10.1016/S2589-7500(22)00091-7.

Tissot, Hegler C. & Lucas A. Pedebos (Dec. 1, 2021). "Improving risk assessment of miscarriage during pregnancy with knowledge graph embeddings". In: *Journal of Healthcare Informatics Research* 5.4, pp. 359–381. DOI: 10.1007/s41666-021-00096-6.

Tong, Catherine et al. (2022). "Predicting patient outcomes with graph representation learning". In: *AI for Disease Surveillance and Pandemic Intelligence: Intelligent Disease Detection in Action*. Ed. by Arash Shaban-Nejad, Martin Michalowski & Simone Bianco. Studies in Computational Intelligence. Cham, CH: Springer International, pp. 281–293. DOI: 10.1007/978-3-030-93080-6_20.

TVCA (Apr. 2021). *Rapid Diagnostic Service Plan 2020–2024*. Thames Valley Cancer Alliance. URL: https://thamesvalleycanceralliance.nhs.uk/rapid-diagnostic-service-plan/ (retrieved 11/26/2024).

— (Sept. 2023). *Lower GI Urgent Suspected Cancer Pathway*. Thames Valley Cancer Alliance. URL: https://thamesvalleycanceralliance.nhs.uk/lower-gi-urgent-pathway/ (retrieved 11/26/2024).

van Dongen, Stijn (May 29, 2000). "Graph clustering by flow simulation". Doctoral thesis. Universiteit Utrecht. URL: https://dspace.library.uu.nl/handle/1874/848 (retrieved 09/09/2024).

— (Jan. 2008). "Graph clustering via a discrete uncoupling process". In: *SIAM Journal on Matrix Analysis and Applications* 30.1, pp. 121–141. DOI: 10.1137/040608635.

van Smeden, Maarten, Frank E. Harrell & Darren L. Dahly (June 1, 2018). "Novel diabetes subgroups". In: *The Lancet Diabetes & Endocrinology* 6.6, pp. 439–440. DOI: 10.1016/S2213-8587(18)30124-4.

van Zelm, Ruben et al. (Sept. 2018). "Protocol for process evaluation of evidence-based care pathways: the case of colorectal cancer surgery". In: *JBI Evidence Implementation* 16.3, p. 145. DOI: 10.1097/XEB.0000000000000149.

van de Klundert, Joris, Pascal Gorissen & Stef Zeemering (Dec. 1, 2010). "Measuring clinical pathway adherence". In: *Journal of Biomedical Informatics* 43.6, pp. 861–872. DOI: 10.1016/j.jbi.2010.08.002.

van der Aalst, Wil M. P. (Aug. 1, 2012). "Process mining". In: *Communications of the ACM* 55.8, pp. 76–83. DOI: 10.1145/2240236.2240257.

— (2016). *Process Mining: Data Science in Action*. 2nd ed. Berlin & Heidelberg, DE: Springer. DOI: 10.1007/978-3-662-49851-4.

— (Jan. 2023). "Object-centric process mining: unraveling the fabric of real processes". In: *Mathematics* 11.12, p. 2691. DOI: 10.3390/math11122691.

Villalobos, Pablo et al. (June 4, 2024). *Will we run out of data? Limits of LLM scaling based on human-generated data*. DOI: 10.48550/arXiv.2211.04325. Pre-published.

Vlietstra, Wytze J. et al. (July 1, 2017). "Automated extraction of potential migraine biomarkers using a semantic graph". In: *Journal of Biomedical Informatics* 71, pp. 178–189. DOI: 10.1016/j.jbi.2017.05.018.

Vlietstra, Wytze J. et al. (Aug. 20, 2020). "Identifying disease trajectories with predicate information from a knowledge graph". In: *Journal of Biomedical Semantics* 11.1, p. 9. DOI: 10.1186/s13326-020-00228-8.

Vogt, Verena, Stefan M. Scholz & Leonie Sundmacher (Apr. 1, 2018). "Applying sequence clustering techniques to explore practice-based ambulatory care pathways in insurance claims data". In: *European Journal of Public Health* 28.2, pp. 214–219. DOI: 10.1093/eurpub/ckx169.

Walker, A. Sarah et al. (July 1, 2017). "Mortality risks associated with emergency admissions during weekends and public holidays: an analysis of electronic health records". In: *The Lancet* 390.10089, pp. 62–72. DOI: 10.1016/S0140-6736(17)30782-1.

Wang, Quan et al. (Dec. 2017). "Knowledge graph embedding: a survey of approaches and applications". In: *IEEE Transactions on Knowledge and Data Engineering* 29.12, pp. 2724–2743. DOI: 10.1109/TKDE.2017.2754499.

Watson, Jessica et al. (Nov. 1, 2017). "Identifying clinical features in primary care electronic health record studies: methods for codelist development". In: *BMJ Open* 7.11, e019637. DOI: 10.1136/bmjopen-2017-019637.

Weller, D. et al. (Mar. 2012). "The Aarhus statement: improving design and reporting of studies on early cancer diagnosis". In: *British Journal of Cancer* 106.7, pp. 1262–1267. DOI: 10.1038/bjc.2012.68.

White, Lyndon et al. (Dec. 8, 2015). "How well sentence embeddings capture meaning". In: *Proceedings of the 20th Australasian Document Computing Symposium*. ADCS 2015. New York, NY, USA: Association for Computing Machinery, pp. 1–8. DOI: 10.1145/2838931.2838932.

White, Nicole et al. (Sept. 4, 2023). "Evidence of questionable research practices in clinical prediction models". In: *BMC Medicine* 21.1, p. 339. DOI: 10.1186/s12916-023-03048-6.

Williams, Richard et al. (2014). "Using string metrics to identify patient journeys through care pathways". In: *AMIA Annual Symposium Proceedings* 2014, p. 1208. PMID: 25954432.

Williams, Richard et al. (June 1, 2017). "Clinical code set engineering for reusing EHR data for research: A review". In: *Journal of Biomedical Informatics* 70, pp. 1–13. DOI: 10.1016/j.jbi.2017.04.010.

Wohed, P. et al. (2006). "On the suitability of BPMN for business process modelling". In: *Business Process Management*. BPM 2006. Ed. by Schahram Dustdar, José Luiz Fiadeiro & Amit P. Sheth. Vol. 4102. Lecture Notes in Computer Science. Berlin & Heidelberg, DE: Springer, pp. 161–176. DOI: 10.1007/11841760_12.

Wornow, Michael et al. (July 29, 2023). "The shaky foundations of large language models and foundation models for electronic health records". In: *npj Digital Medicine* 6.1 (1), pp. 1–10. DOI: 10.1038/s41746-023-00879-8.

Wyatt, Jeremy C. & David Spiegelhalter (1991). "Field trials of medical decision-aids: potential problems and solutions." In: *Proceedings of the Annual Symposium on Computer Application in Medical Care*, pp. 3–7. PMID: 1807610.

Wyatt, Jeremy C. et al. (2023). "Which computable biomedical knowledge objects will be regulated? Results of a UK workshop discussing the regulation of knowledge libraries and software as a medical device". In: *Learning Health Systems* 7.4, e10386. DOI: 10.1002/lrh2.10386.

Xu, Zhongzhi et al. (Oct. 2019). "Explainable learning for disease risk prediction based on comorbidity networks". In: *IEEE International Conference on Systems, Man and Cybernetics*. SMC 2019, pp. 814–818. DOI: 10.1109/SMC.2019.8914644.

Yang, Bishan et al. (May 9, 2015). "Embedding entities and relations for learning and inference in knowledge bases". In: *3rd International Conference on Learning Representations*. ICLR 2015. Ed. by Yoshua Bengio & Yann LeCun. San Diego, CA, USA. DOI: 10.48550/arXiv.1412.6575.

Yang, Wei & Qiang Su (June 2014). "Process mining for clinical pathway: literature review and future directions". In: *11th International Conference on Service Systems and Service Management*. ICSSM 2014. Beijing, CN, pp. 1–5. DOI: 10.1109/ICSSSM.2014.6943412.

Yoon, Byoung-Ha, Seon-Kyu Kim & Seon-Young Kim (Mar. 2017). "Use of graph database for the integration of heterogeneous biological data". In: *Genomics & Informatics* 15.1, pp. 19–27. DOI: 10.5808/GI.2017.15.1.19.

Zhang, Joe et al. (Oct. 1, 2023). "Mapping and evaluating national data flows: transparency, privacy, and guiding infrastructural transformation". In: *The Lancet Digital Health* 5.10, e737–e748. DOI: 10.1016/S2589-7500(23)00157-7.

Zou, Will Y. et al. (Oct. 2013). "Bilingual word embeddings for phrase-based machine translation". In: *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*. EMNLP 2013. Ed. by David Yarowsky et al. Seattle, WA, USA: Association for Computational Linguistics, pp. 1393–1398. URL: https://aclanthology.org/D13-1141 (retrieved 04/09/2024).

zur Muehlen, Michael & Jan Recker (2008). "How much language is enough? Theoretical and practical use of the Business Process Modeling Notation". In: *Advanced Information Systems Engineering*. CAiSE 2008. Ed. by Zohra Bellahsène & Michel Léonard. Vol. 5074. Lecture Notes in Computer Science. Montpellier, FR: Springer, pp. 465–479. DOI: 10.1007/978-3-540-69534-9_35.

A *Embedding model evaluation*

This appendix accompanies Section 4.2.3, and expands on the evaluation of each embedding model. Each figure in this section summarises each knowledge graph embedding model, following the method used by Fu et al. (2023). Histograms show the distribution of values for both the embedding-based distance and OPCS tree distance. Heatmaps show the average distance between concepts in each chapter, and boxplots summarise the average distance between concepts in the same and in different chapters. Finally, scatter plots show the correlation between the embedding distance and OPCS tree distance, both overall and split by OPCS chapter.

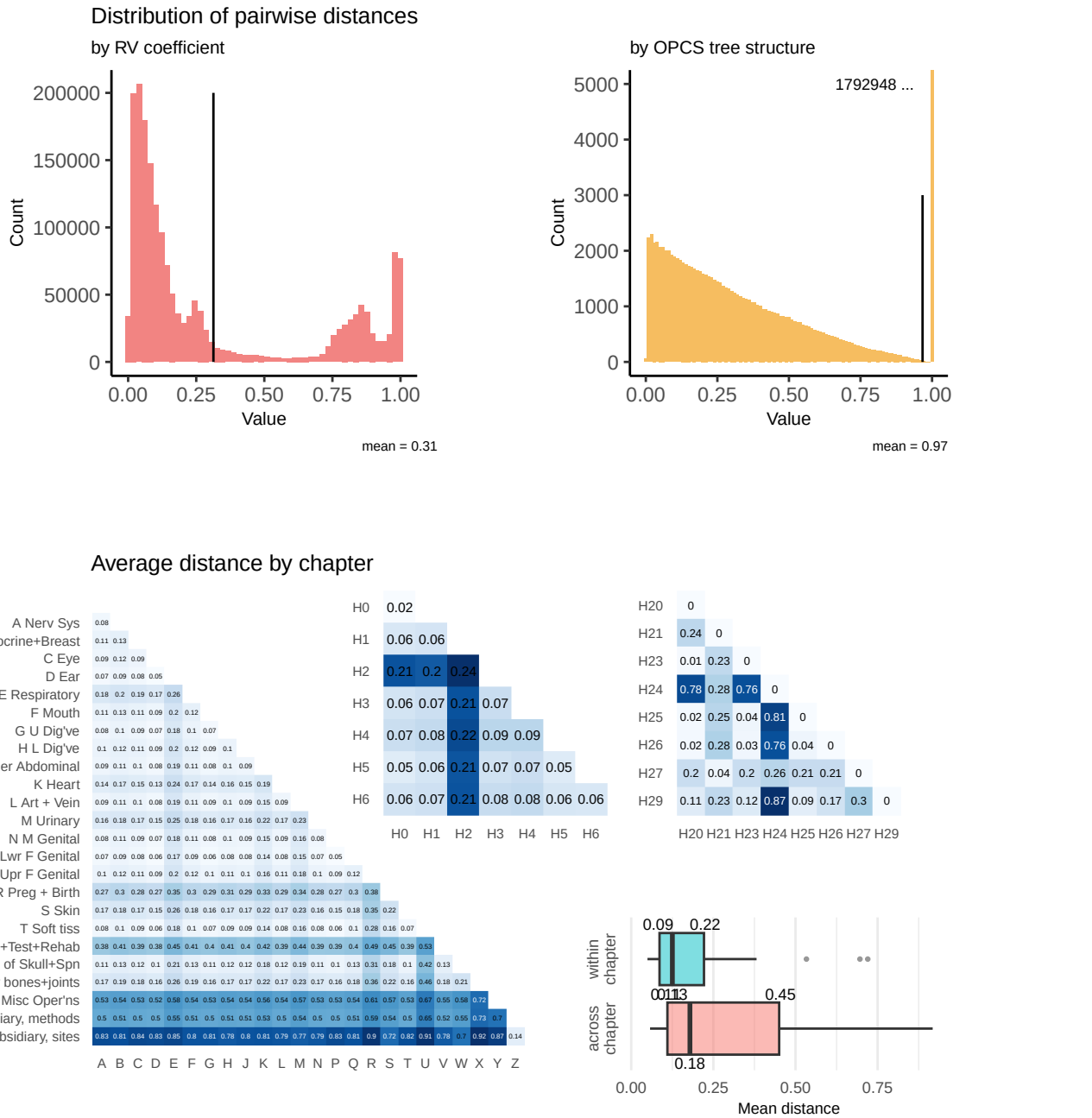

Figure A.1. Summary statistics for the TransE model

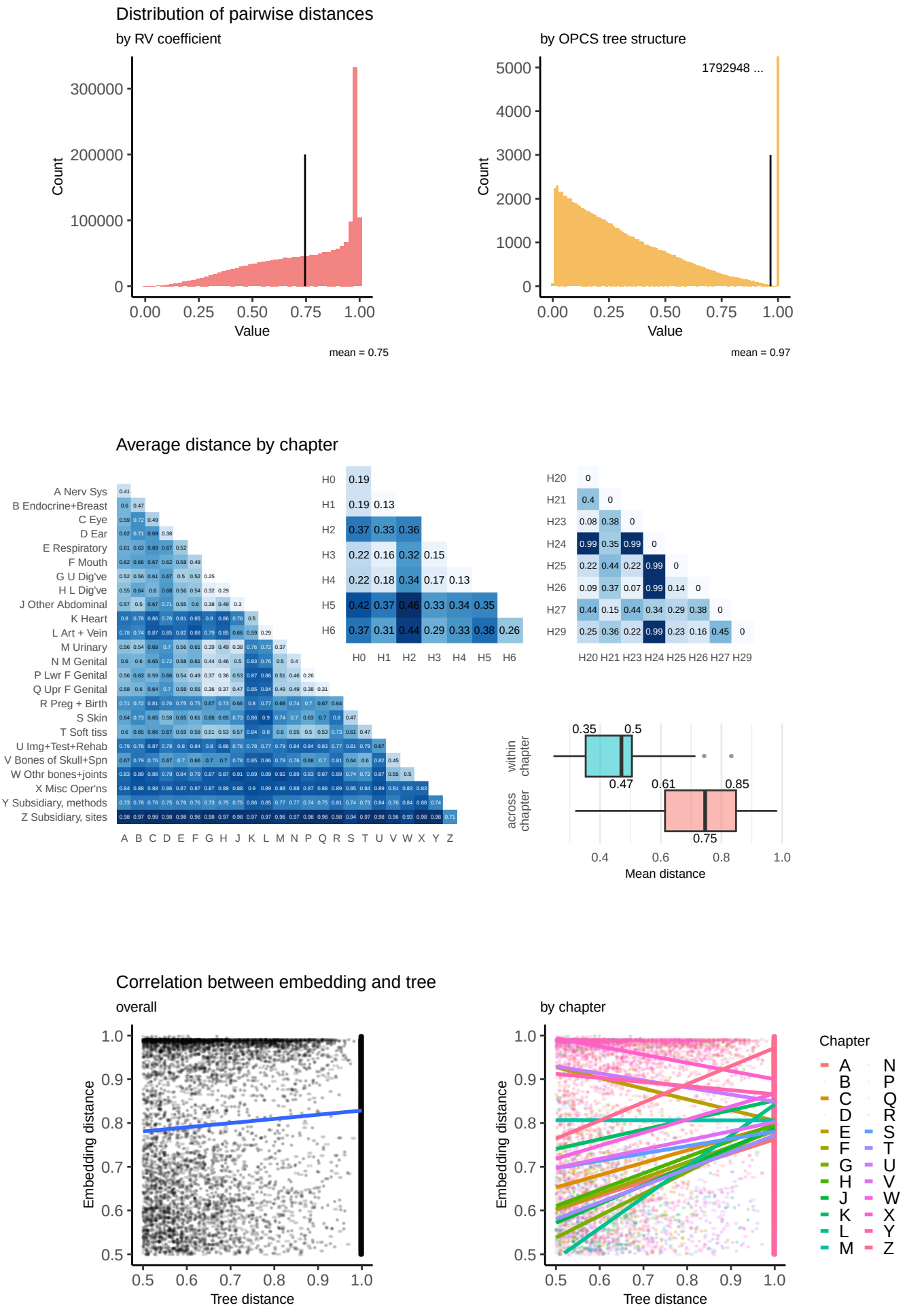

Figure A.2. Summary statistics for the DistMult model

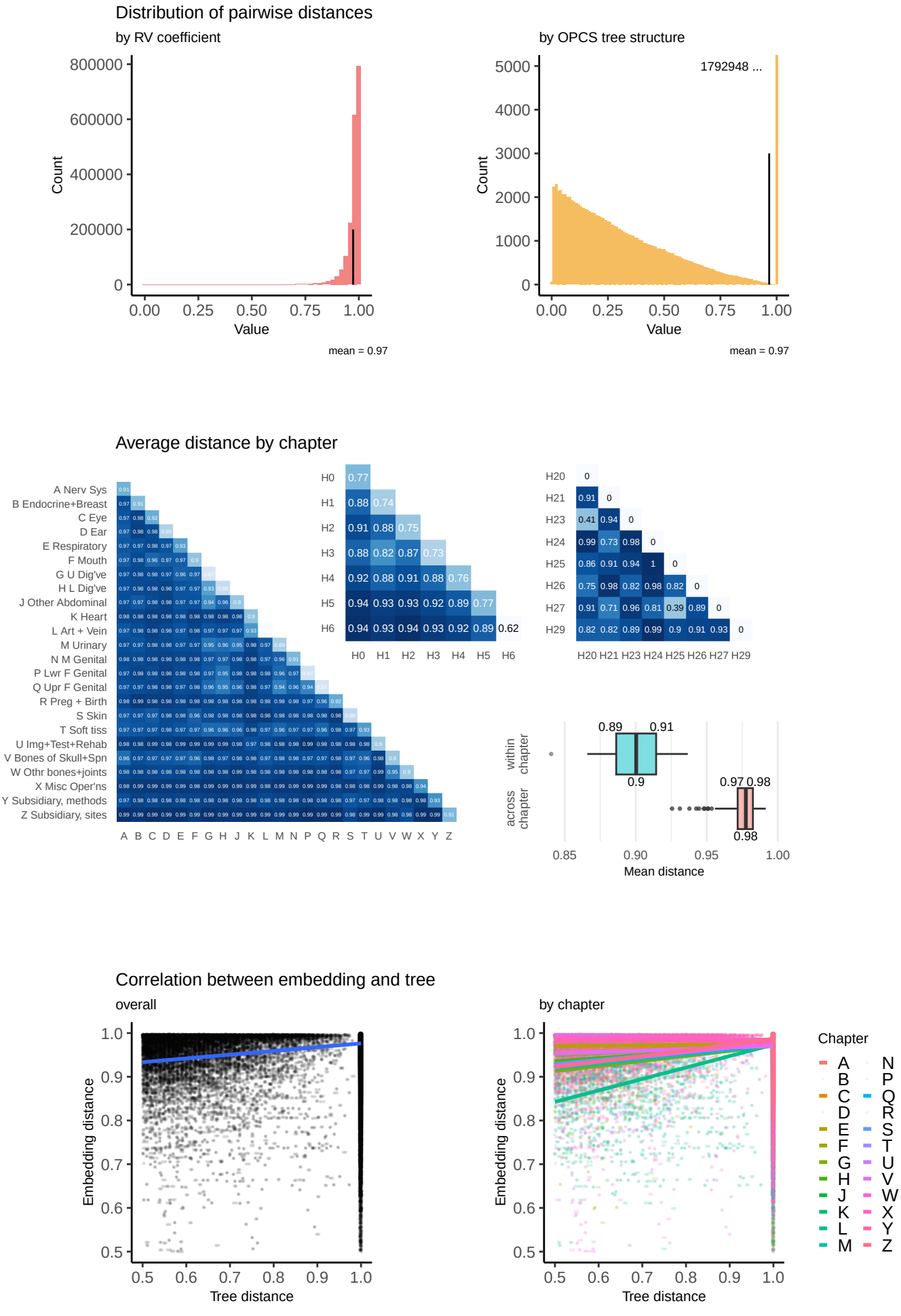

Figure A.3. Summary statistics for the RotatE model

B *Comparison of distance metrics*

This appendix accompanies Section 4.3.2, and provides a sample of results from each distance metric. Each figure in this appendix shows the timelines of 32 patients randomly sampled from the OUH dataset (8 each of length 1, 2, 4, and 8), followed by their 20 nearest neighbours according to the respective distance metric. Basic composition methods are listed on pages 170–172, sequence-based methods on pages 173–175, and E-DTW variants on pages 176–183.

Figure B.1. Key to OPCS concepts used in Figures B.2 – B.15

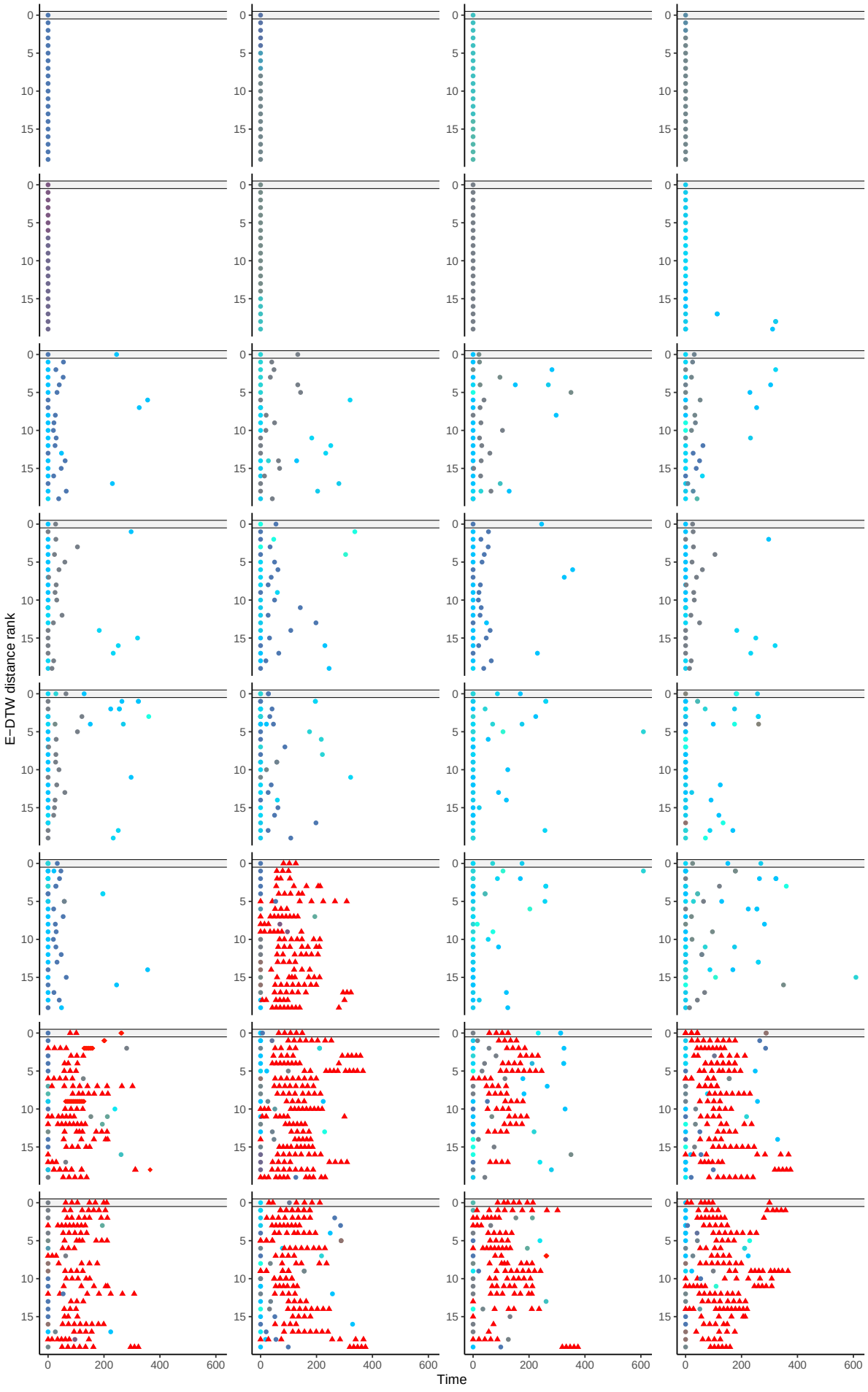

Figure B.2. Nearest neighbours for 32 randomly sampled pathways, according to the cosine distance between product of embeddings

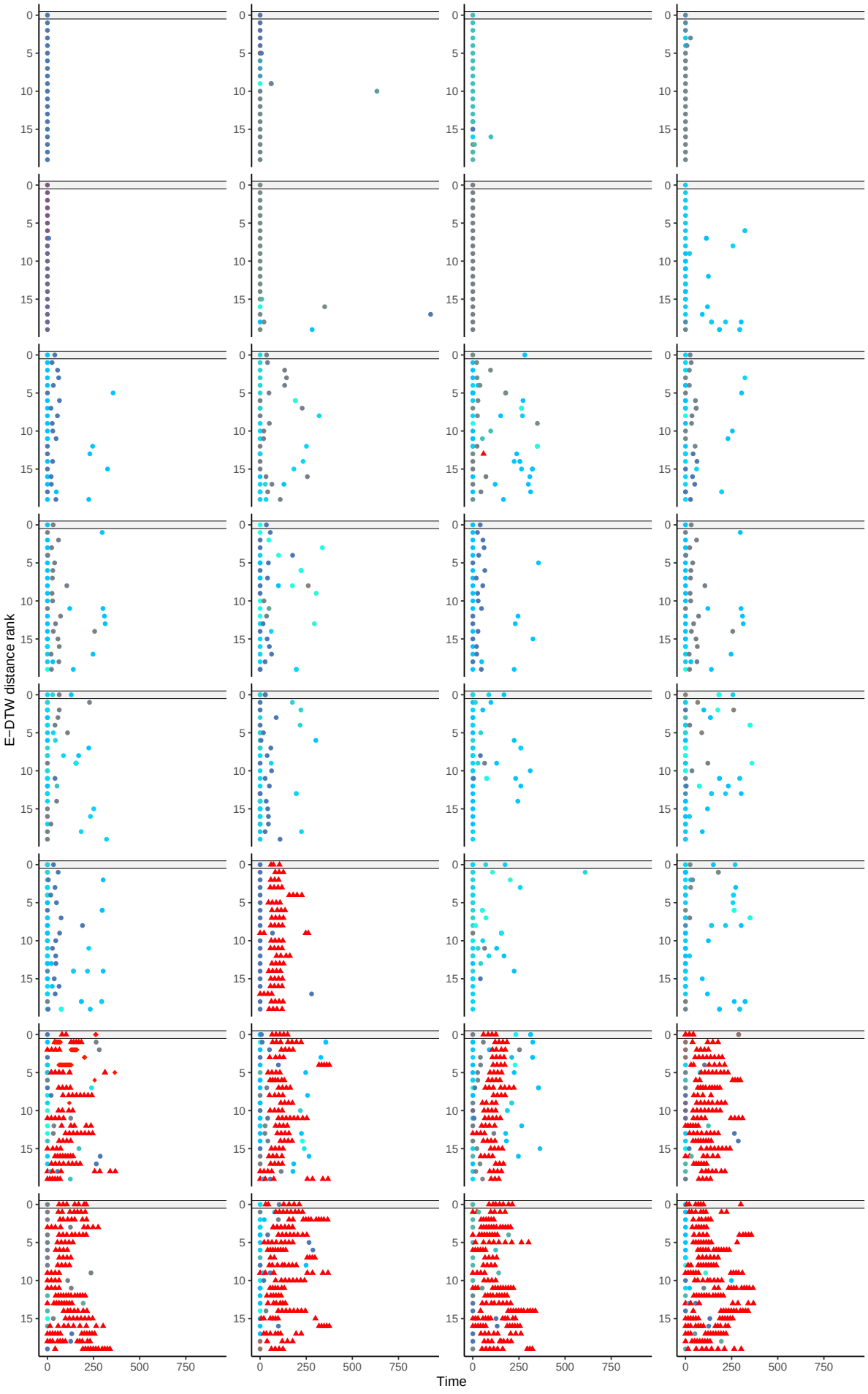

Figure B.3. Nearest neighbours for 32 randomly sampled pathways, according to the cosine distance between mean of embeddings

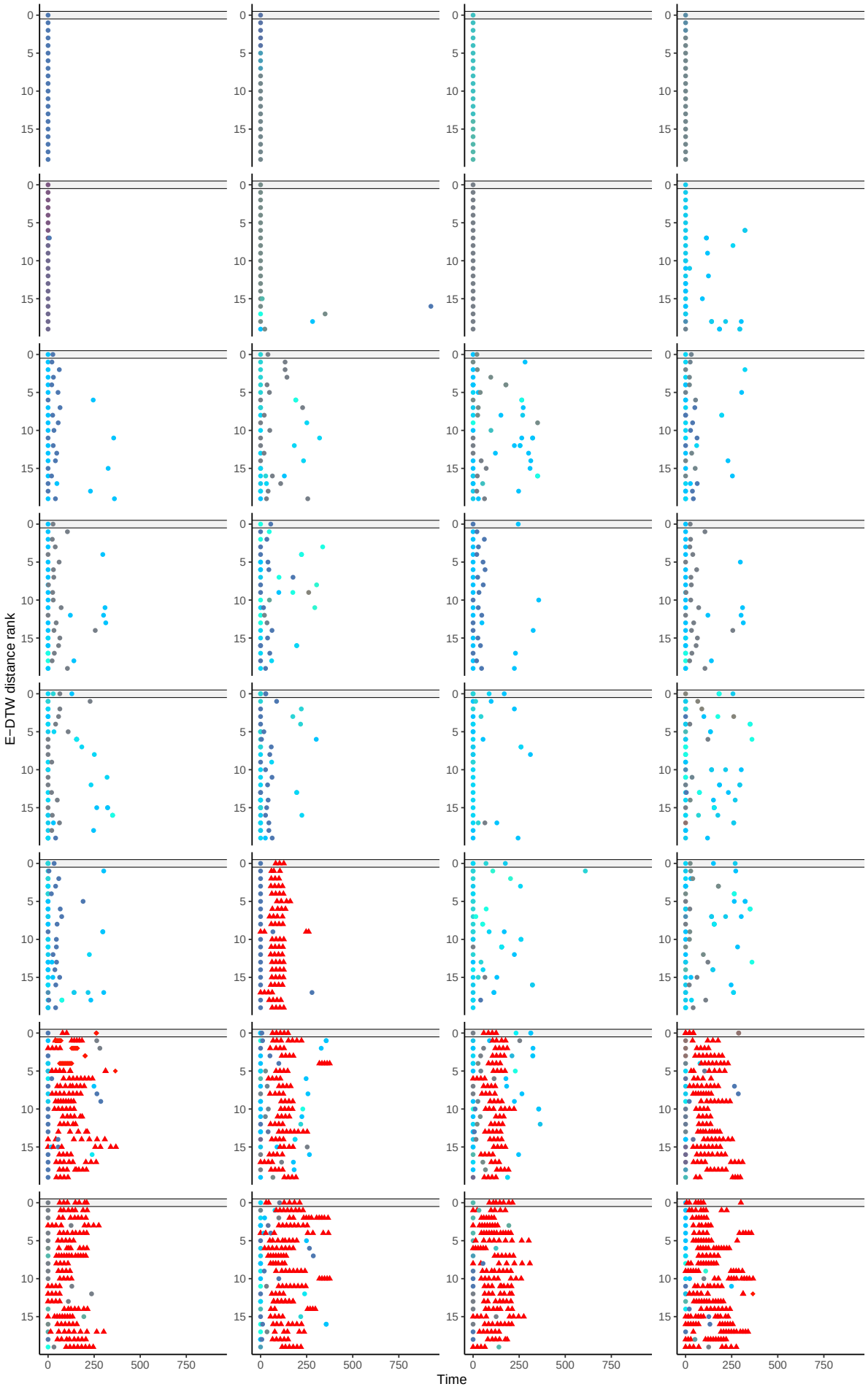

Figure B.4. Nearest neighbours for 32 randomly sampled pathways, according to the cosine distance between mean + variance of embeddings

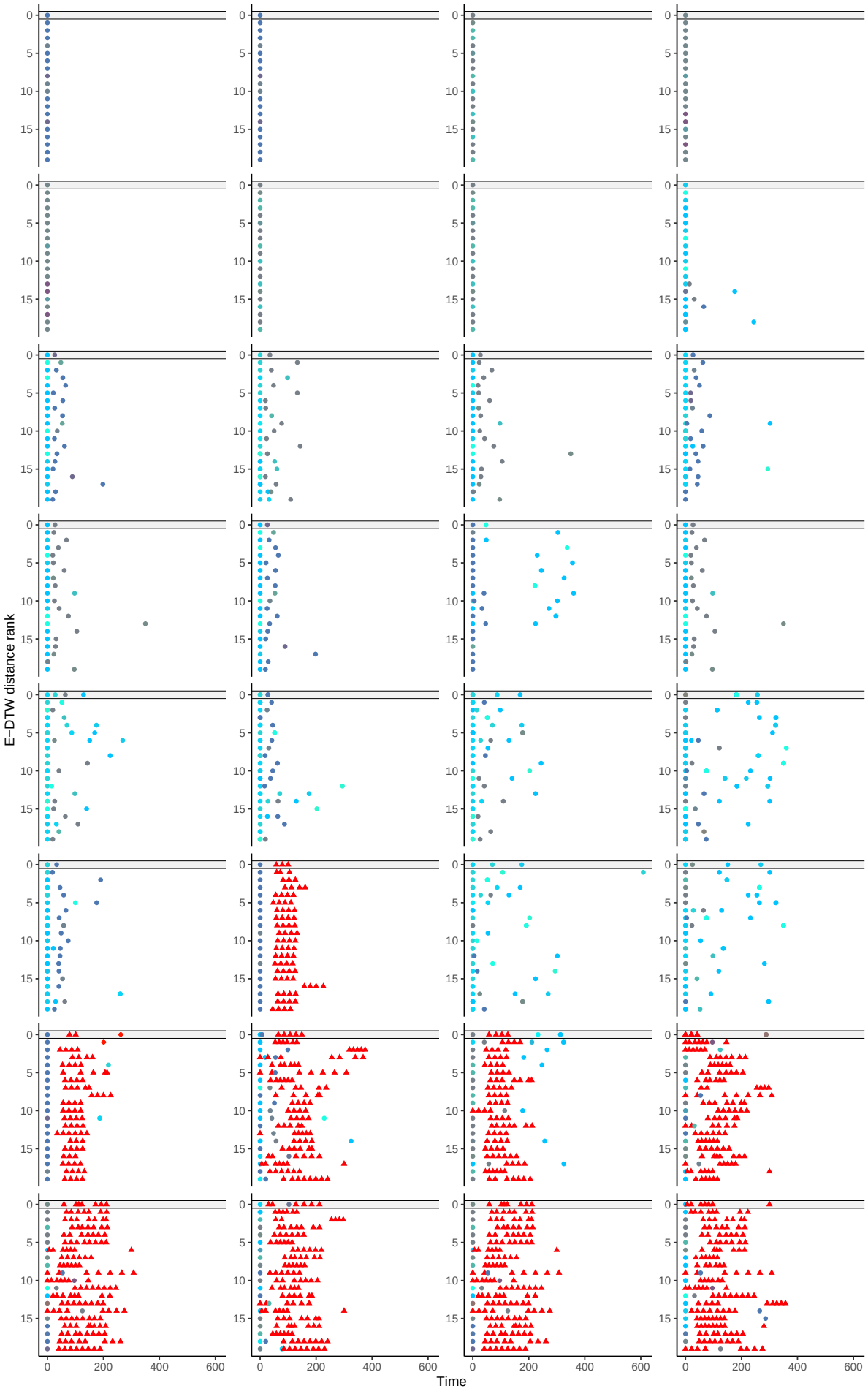

Figure B.5. Nearest neighbours for 32 randomly sampled pathways, according to the Needleman-Wunsch algorithm

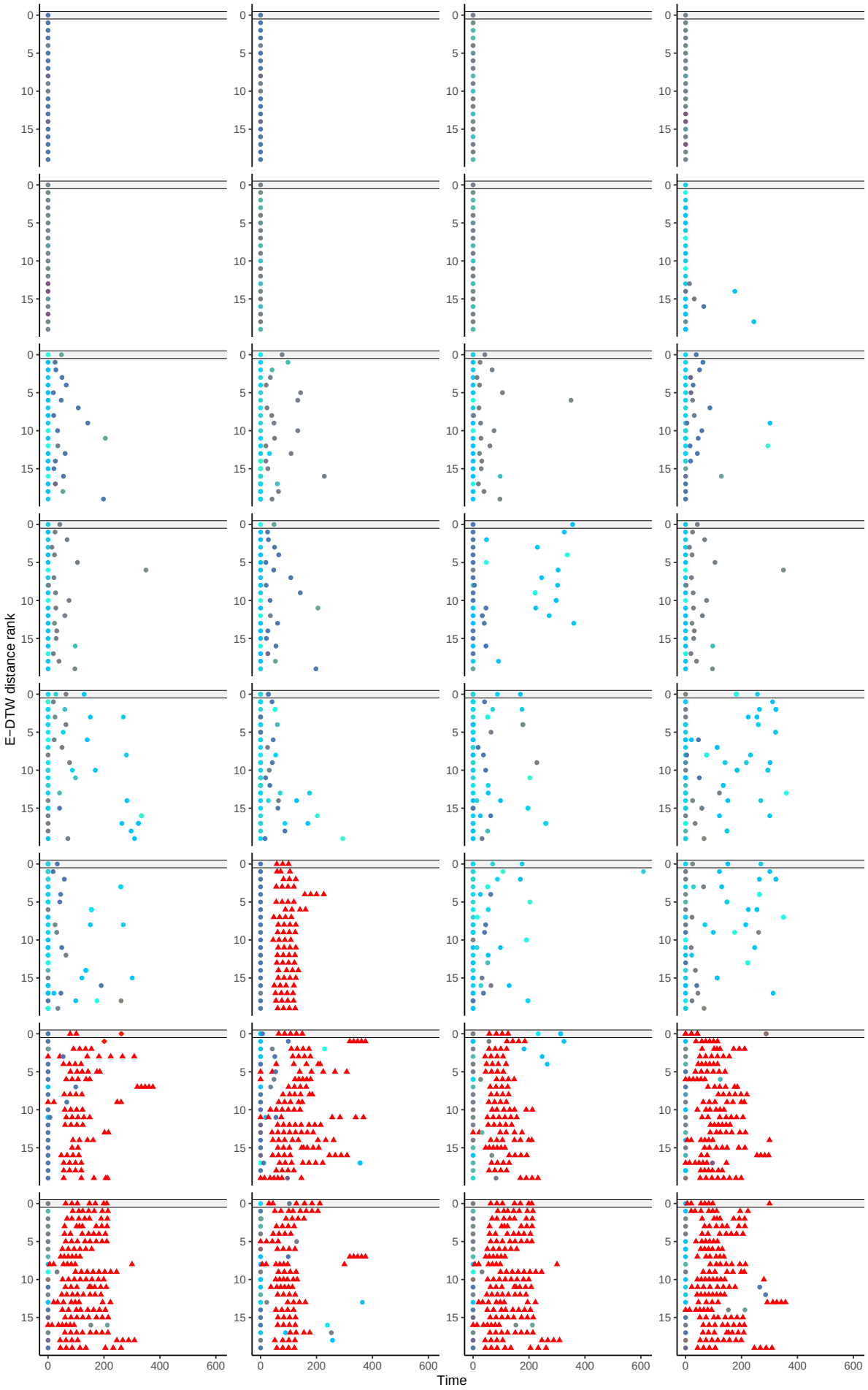

Figure B.6. Nearest neighbours for 32 randomly sampled pathways, according to the longest common subsequence

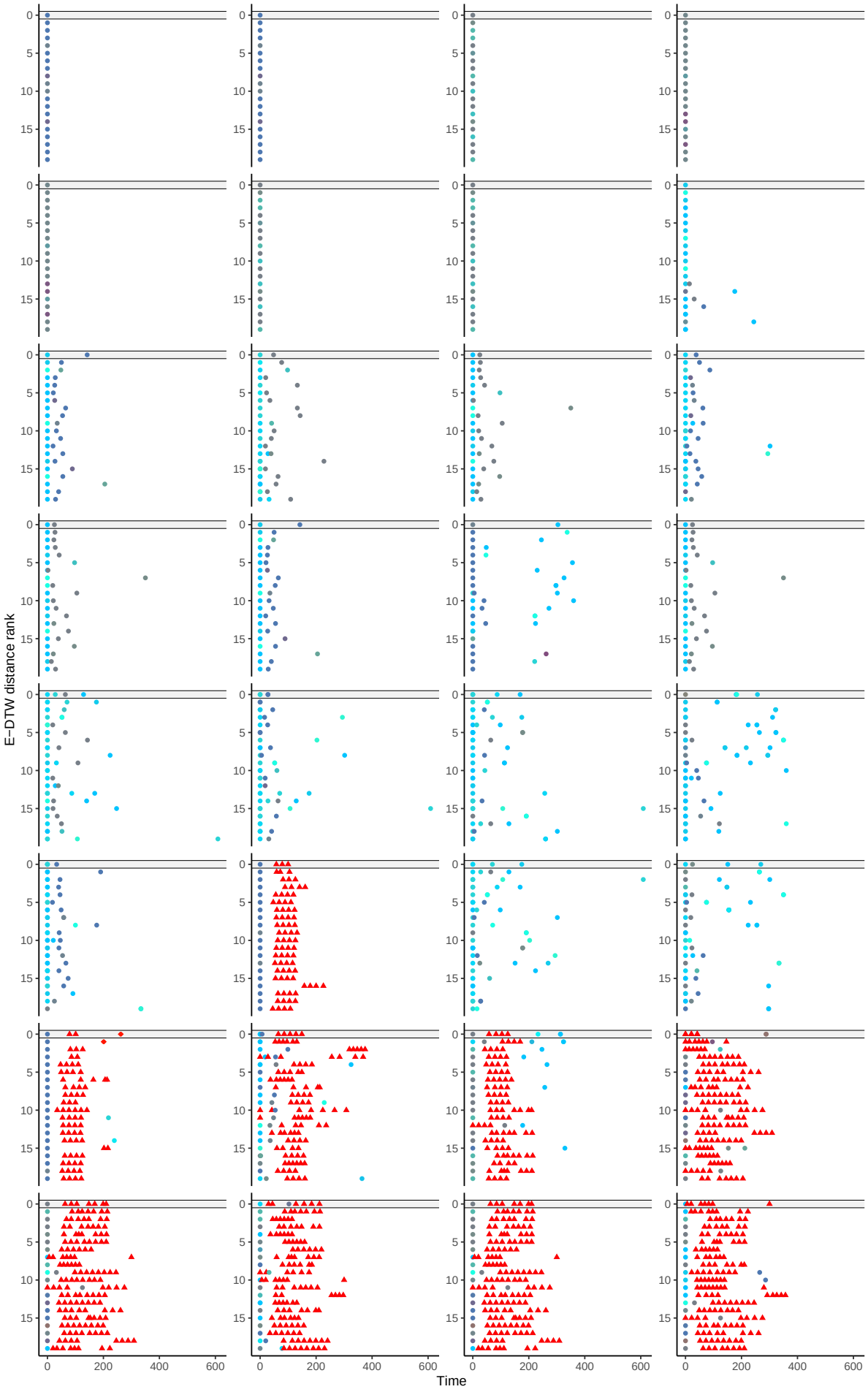

Figure B.7. Nearest neighbours for 32 randomly sampled pathways, according to the Damerau-Levenshtein distance

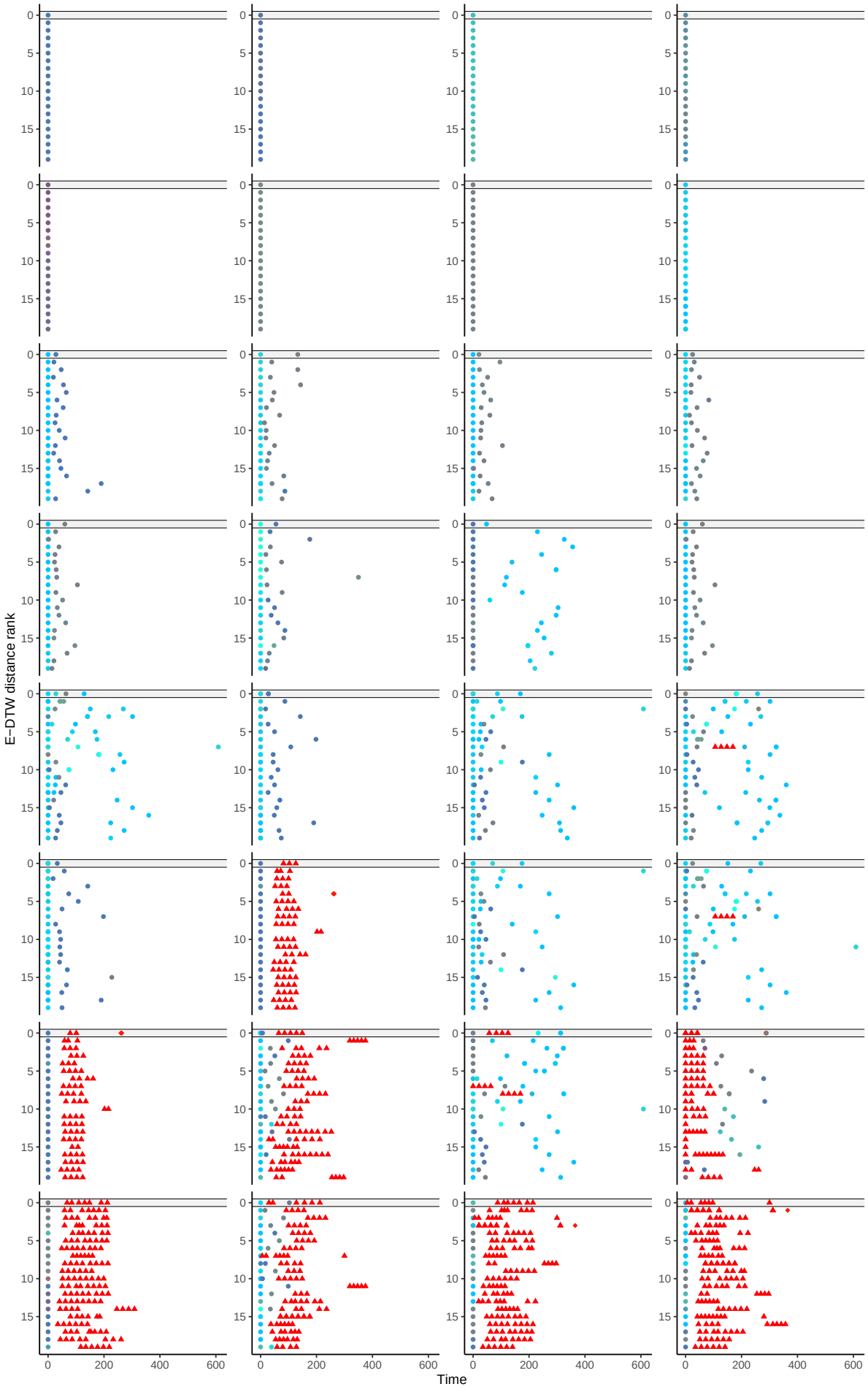

Figure B.8. Nearest neighbours for 32 randomly sampled pathways, according to the E-DTW $(2d, 0)$ measure

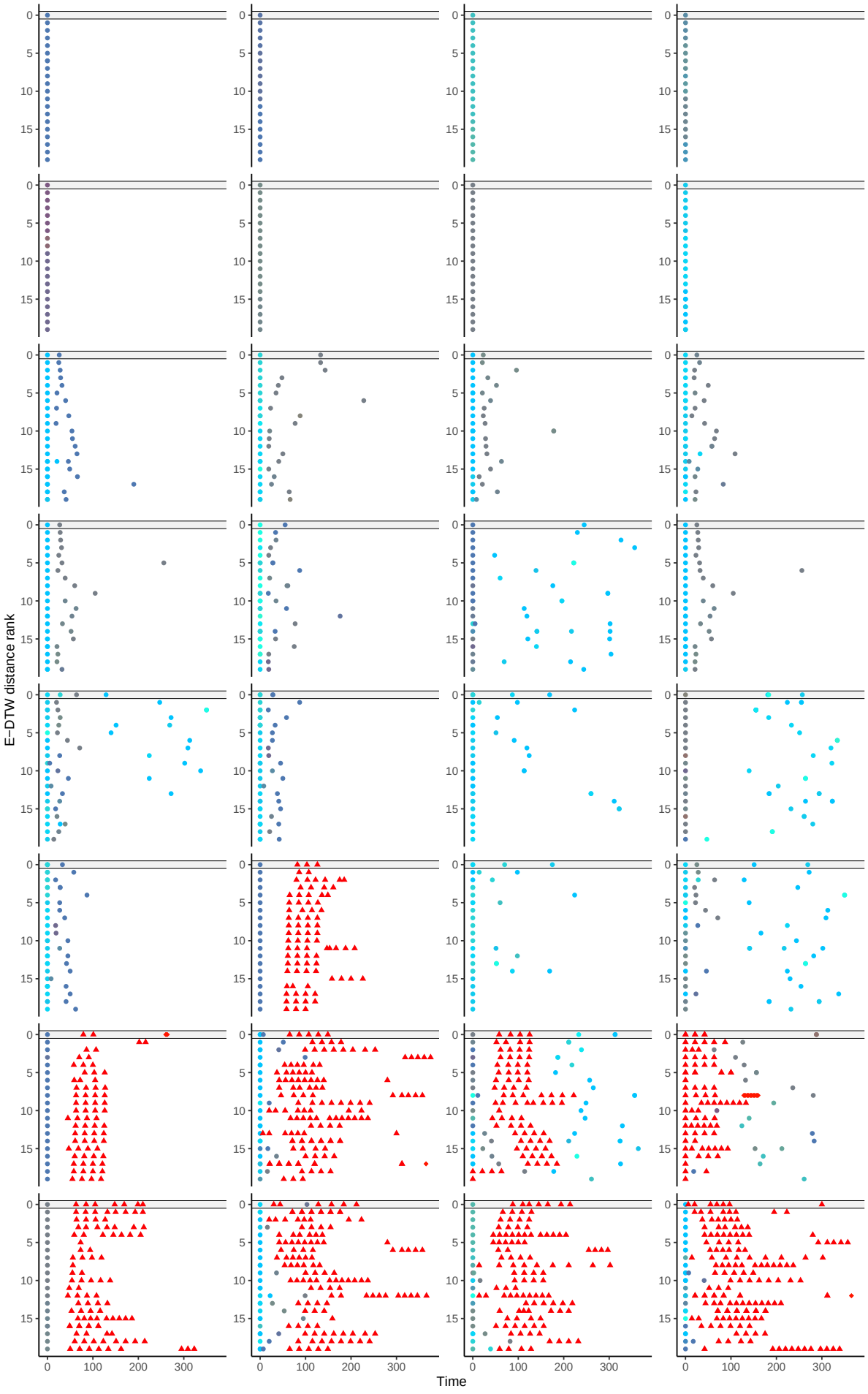

Figure B.9. Nearest neighbours for 32 randomly sampled pathways, according to the E-DTW $(2d, i)$ measure

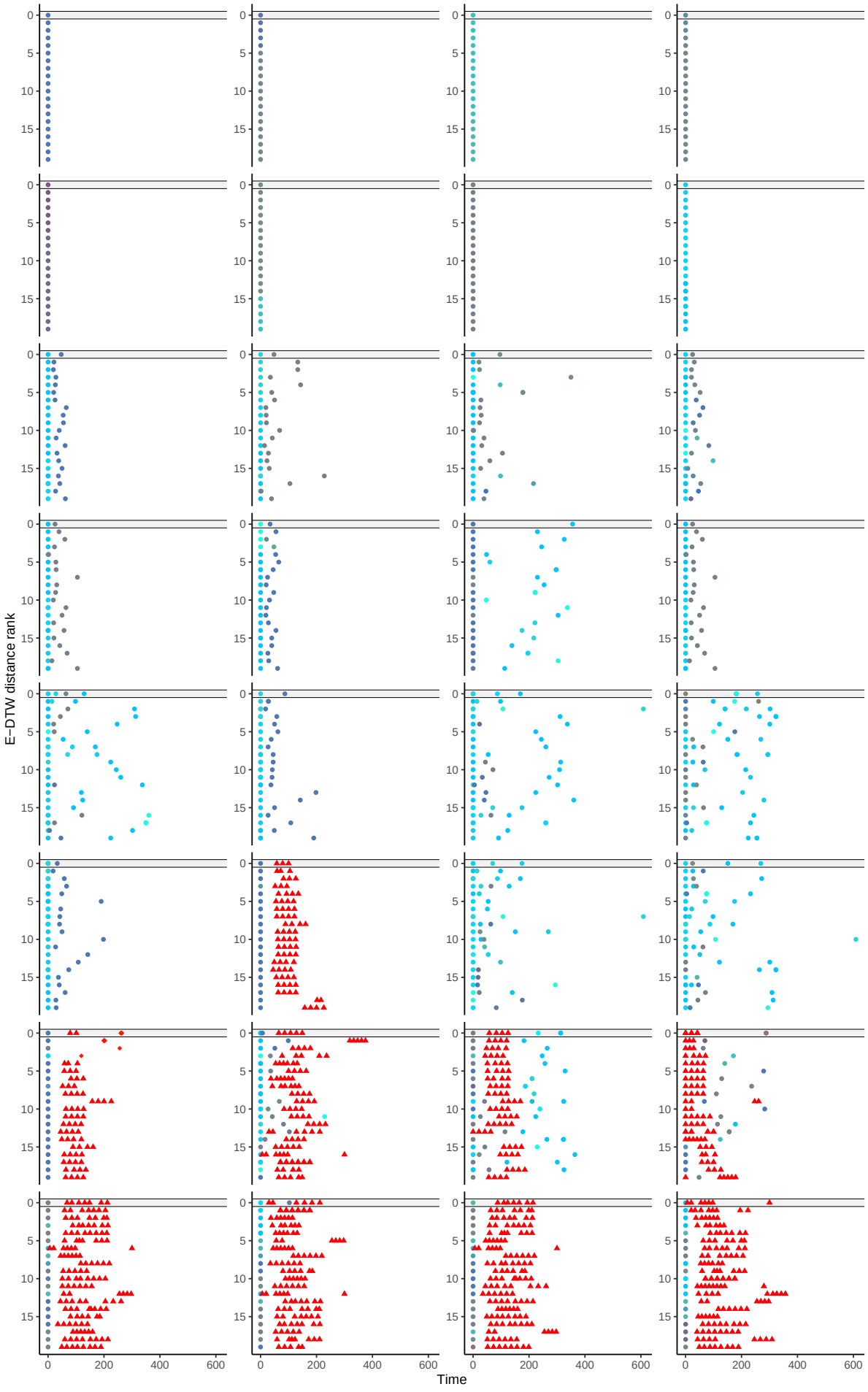

Figure B.10. Nearest neighbours for 32 randomly sampled pathways, according to the E-DTW $(nd, 0)$ measure

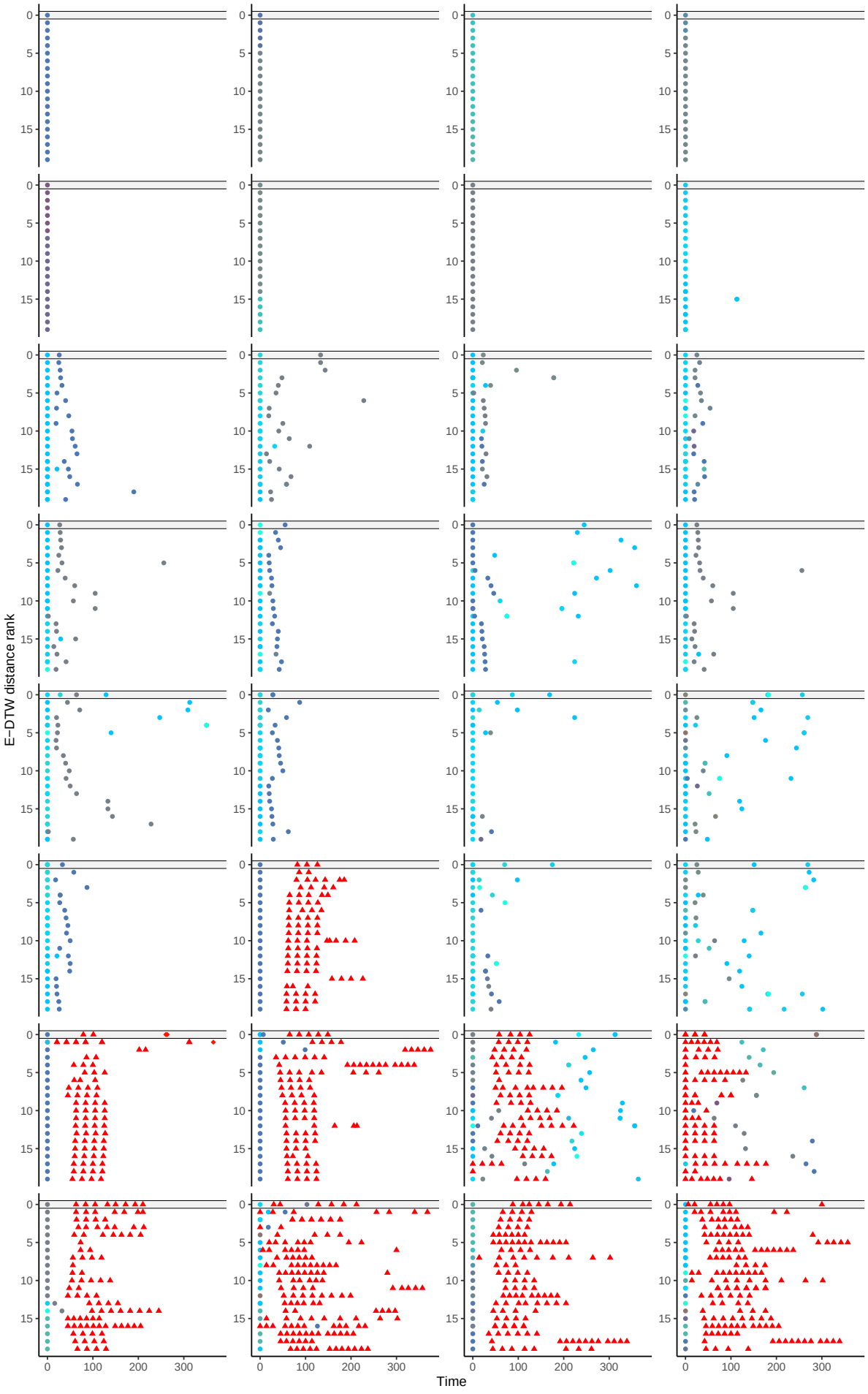

Figure B.11. Nearest neighbours for 32 randomly sampled pathways, according to the E-DTW (nd, i) measure

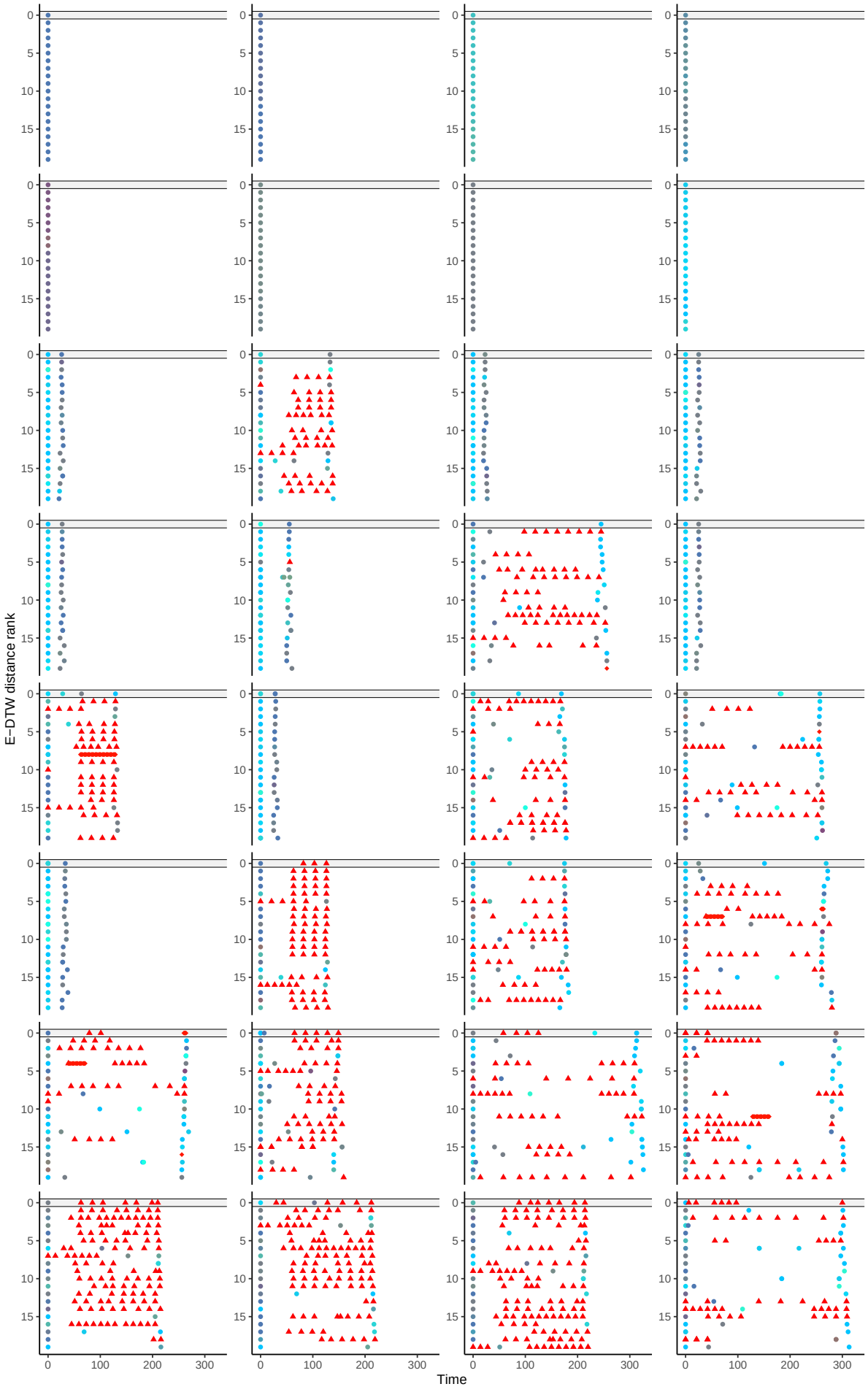

Figure B.12. Nearest neighbours for 32 randomly sampled pathways, according to the E-DTW $(2d, 0) + p$ measure

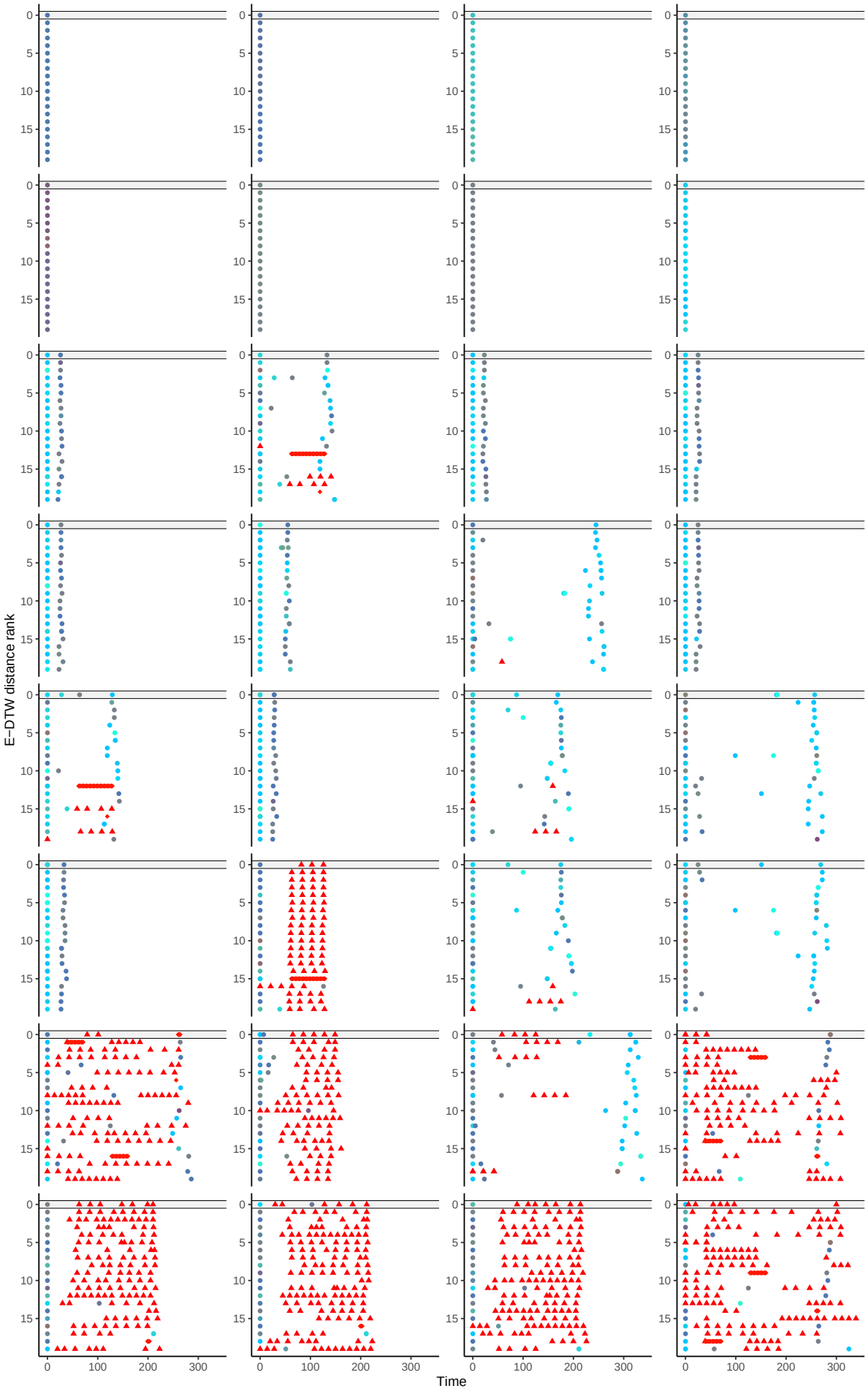

Figure B.13. Nearest neighbours for 32 randomly sampled pathways, according to the E-DTW $(2d, i) + p$ measure

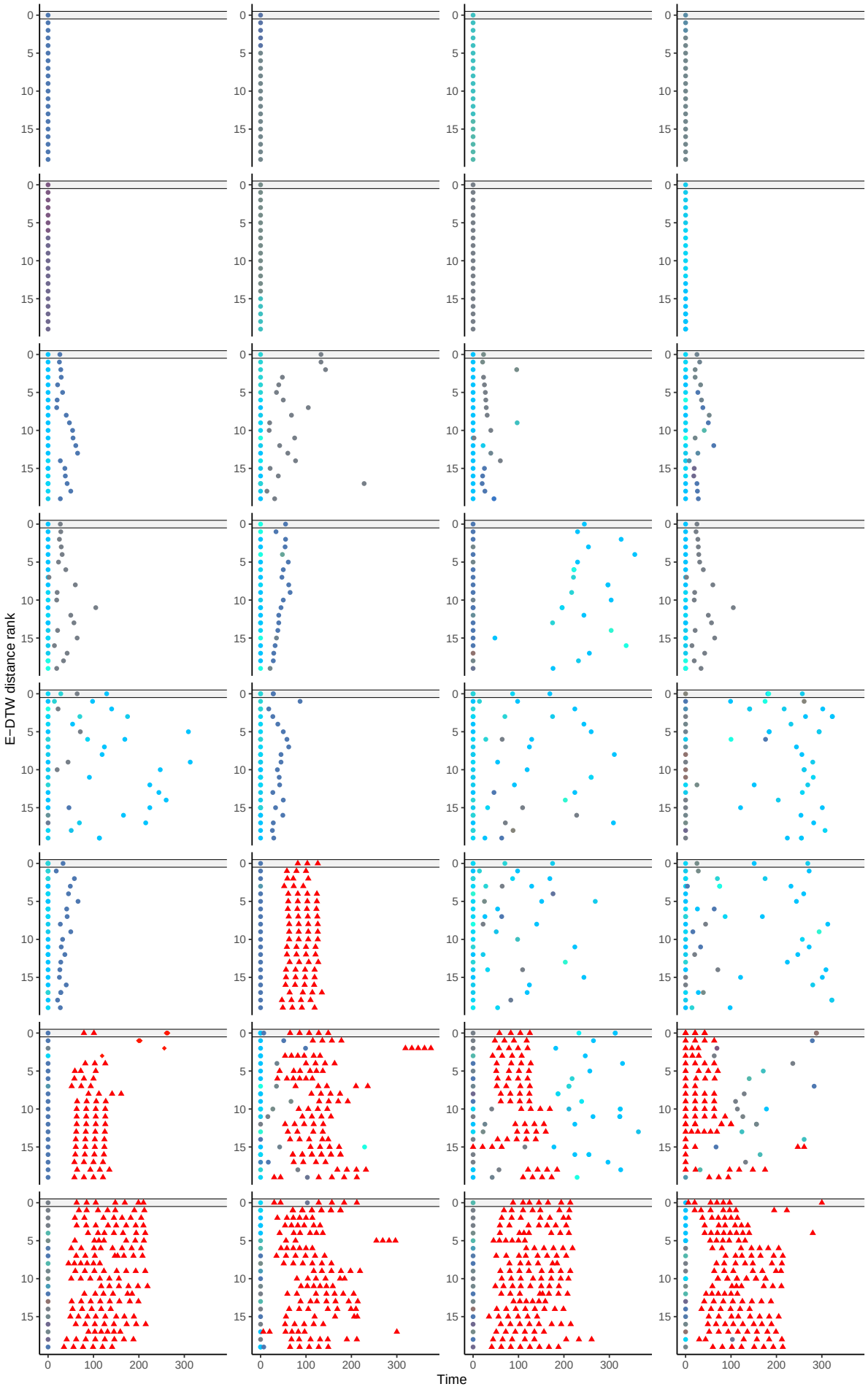

Figure B.14. Nearest neighbours for 32 randomly sampled pathways, according to the E-DTW $(nd, 0) + p$ measure

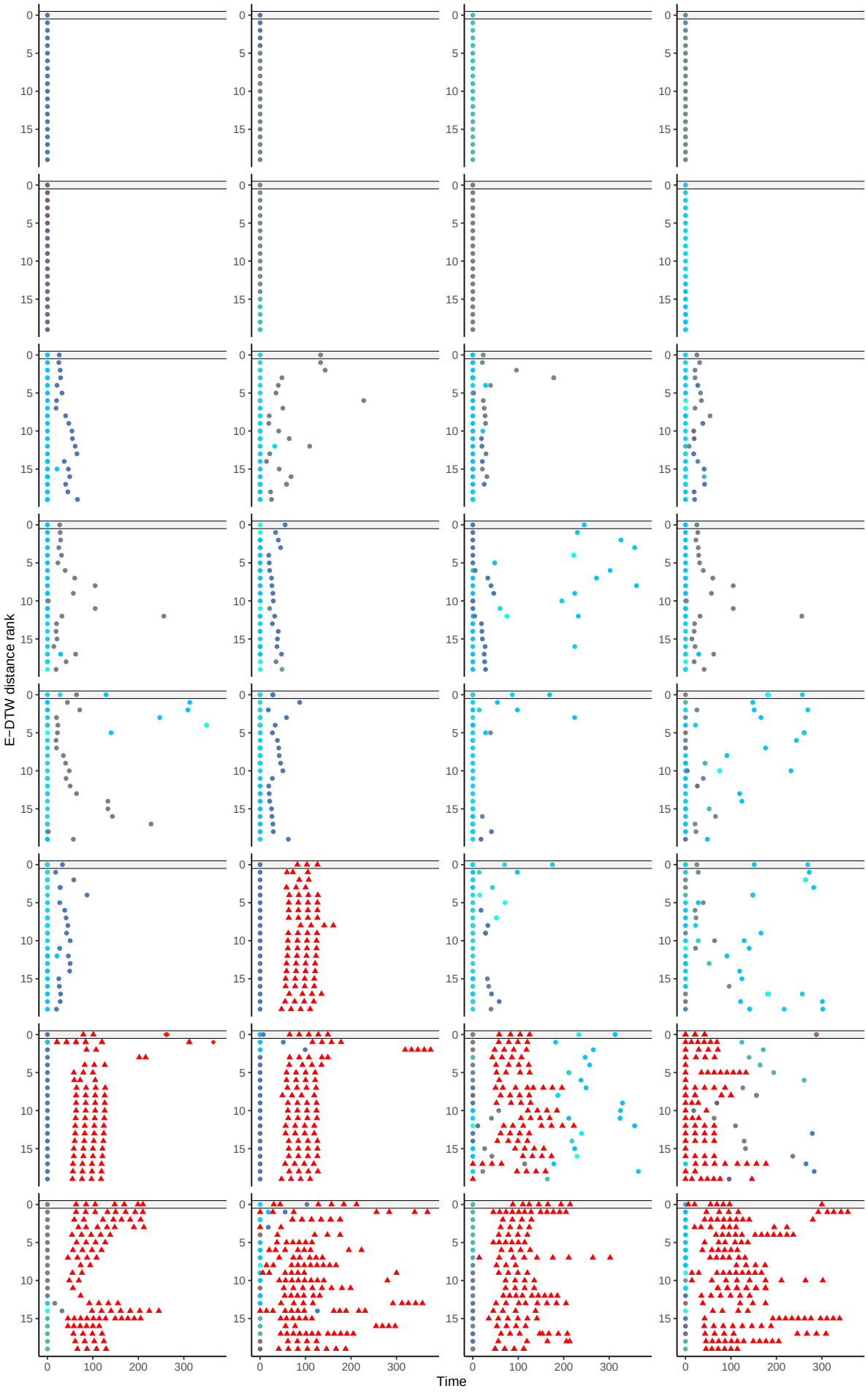

Figure B.15. Nearest neighbours for 32 randomly sampled pathways, according to the E-DTW $(nd, i) + p$ measure